CLASSICAL AND BAYESIAN INSTRUMENT DEVELOPMENT

By

Lili Garrard

Submitted to the graduate degree program in Biostatistics and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy

_____
Chairperson Byron J. Gajewski, Ph.D.

_____
Marjorie J. Bott, Ph.D., RN

_____
Jianghua He, Ph.D.

_____
Jo A. Wick, Ph.D.

_____
Hung-Wen Yeh, Ph.D.

Date Defended: December 18, 2015

The dissertation Committee for Lili Garrard
certifies that this is the approved version of the following dissertation:

CLASSICAL AND BAYESIAN INSTRUMENT DEVELOPMENT

_____
Chairperson Byron J. Gajewski, Ph.D.

Date approved: December 21, 2015

Abstract

Both patient-reported outcome measures (PROMs) and clinician-reported outcome (ClinRO) measures are recognized as essential tools for advocating patient-centered care, an important driving force behind the current U.S. health care system. Close collaborations among the research community and regulatory bodies have been initiated to form standardized guidelines for the development and evaluation of PROMs and many ClinRO measures that often are designed as psychometric instruments with ordinal response scales. Classical (i.e., frequentist) instrument development often is time-consuming and challenged by small samples (e.g., cases of rare diseases). An innovative Ordinal Bayesian Instrument Development (OBID) approach within a Bayesian Item Response Theory (IRT) framework is introduced to overcome both small sample size and ordinal data modeling challenges, through efficient integration of content validity and construct validity analyses. The performance of OBID is evaluated under a simulation setting with three different types of expert bias (i.e., unbiased, moderately biased, and highly biased), and further evaluated with an exact Bayesian leave-one-out cross-validation (LOO-CV) approach using real data applications. Results successfully demonstrated the OBID approach as a promising tool in future PROMs and ClinRO measures development for small populations or rare diseases. Alternatively classical psychometric methodologies are efficient and reliable with relatively large sample sizes. This study also presents the classical psychometric evaluation of the National Database of Nursing Quality Indicators® (NDNQI®) falls with injury measure, an essential ClinRO measure that supports health care quality improvement efforts and continuous injurious falls research.

Acknowledgements

Nineteen years ago, I made the biggest decision in my life—moving to the United States to pursue a quality education. Being young and fearless, I never looked back and never realized that it would take 12 years for me to return home again. Since then, I had to make many important decisions, with pursing a Ph.D. degree being the most rewarding one. The journey that has led me to the decision of pursuing a Ph.D. was met with many struggles, and I would not have been able to complete this journey without the love and support of many extraordinary people in my life. I truly am the luckiest of all.

First of all, I extend my sincere gratitude and appreciation to my advisor, Dr. Byron J. Gajewski, for putting the idea of "getting a Ph.D." in my head when I was feeling comfortable in life with a Master's degree and a decent job. Dr. Gajewski and I collaborated on a few projects when we both worked for the National Database of Nursing Quality Indicators® (NDNQI®). Dr. Gajewski, thank you for seeing the potential in me when I had numerous doubts about going back to school after being out for six years. Thank you for helping me make the final decision by reminding me that the application deadline had passed. On this note, I also would like to thank Dr. Michael Brimacombe of Biostatistics for kindly re-opening the on-line application, just for me. Dr. Gajewski truly cares about his students and always has our best interest in heart. Thank you for always being honest and pushing us to strive for the best. The "deadlines" really did work well during those moments of low motivation. Without a doubt, it is through your endless patience, priceless mentoring, inspiring guidance, and countless encouragements that I am able to achieve this great accomplishment.

I would like to thank my wonderful committee members, Dr. Marjorie J. Bott (School of Nursing), Dr. Jianghua (Wendy) He, Dr. Jo A. Wick, and Dr. Hung-Wen (Henry) Yeh, for their insightful suggestions on improving my work and their encouragements during the past few years. Dr. Bott, thank you for your support during my NDNQI days and taking the time to review my work despite your busy schedule as the Dean of SON. Dr. He, thank you for your sincere advices with school, career choices, and life. Dr. Wick, thank you for always being kind and your openness to share the most honest advices. Dr. Yeh, thank you for always asking great questions about my research and your insightful suggestions. I always have enjoyed our casual conversations from time to time, even if it is about sleep insomnia.

My gratitude is also extended to Dr. Larry R. Price (Texas State University) for his intellectual support on my research and the opportunity to co-author publications with him. I want to thank my fellow colleague, Alex Karanevich, for turning my dissertation codes into a nice software and co-authoring the software paper with me. Now more people can benefit from the work in this dissertation. I also would like to thank Dr. Matthew S. Mayo (chair of Biostatistics) for his great vision and initiatives that have made the Biostatistics department a top-notch training ground for all the students. I want to thank The University of Kansas Cancer Center, the American Nurses Association, NDNQI, and Drs. Heather Gibbs and Kimberly Engelman for their support on my research through funding and access to data.

There is a special place in my heart for all the wonderful people that I've worked with at NDNQI. I have become good friends with many of them during my six years of tenure there. Thank you, Dr. Diane Boyle, for your priceless friendship, mentorship, and encouragements when I tried to balance school and full-time work. Thank you for keeping me calm and always caring about me. One of these days I finally will beat you on Words with Friends. I want to thank

Dr. Peggy Miller for taking me under her wing when I was a newbie, always trusting my work, and being proud of me when I began the Ph.D. journey. Thank you, Dr. Nancy Dunton, for giving me the opportunity to strive on your great project and your valuable advices with graduate school. Special heart-felt thanks go to the entire analyst team, the best team I could ever ask for. Thank you all for being extremely understanding and supportive, and extending a helping hand on those insane days with competing deadlines between school and work. Thanks to Drs. Chenjuan (Tina) Ma, Shin Hye Park, and JiSun Choi (in Korea) for your wonderful friendships and tremendous support over these years.

I would like to thank my girls from Maryland: Lin Liu, Kyoko Miki, Ola Onolaja, and Moronke Akintunde. Our friendships began when we sat down at the same lunch table at work nine years ago, and I thank you for all your continuous support and love during the hardest times.

I would like to thank my high school vice principal and godfather, Mr. Mike Maggart, for always being there for me during the past 19 years, especially during those confusing years as a teenager. I hope I make you proud.

I am extremely thankful to my dearest friends from Kennedy High, Michigan Tech and KUMC. I am who I am because of all of you. Special thanks go to Jenny Toha, Santiago Aguilar, Dr. Adan Niu (in China), Dr. Chang Liu, Dr. Lezi E, Beibei Xu, Dr. Fran Nunez, Dr. Wei (Will) Jiang, and Dr. Milan Bimali for your love and support. Thank you for never advising me to quit and simply being there for me in the good and bad times. Janelle Noel, you are an angel, and I treasure our friendship so much. We worked through the peaks and valleys of this Ph.D. journey together, sharing laughter and tears. Thank you for everything. I especially want to thank my wonderful academia sisters (as Byron called us), Dr. Yu (Joyce) Jiang and Yang Lei, for your

tremendous academic and emotional supports. Those late night study sessions over the webcam will become one of my best memories in life.

I would like to extend my heart-felt thanks to my family, especially my mother, Mrs. Kefei Li, for the sacrifices she has made for me, and all the love and patience she has shown throughout my life. I want to thank my step-father, Mr. Weiming Li, for simply loving my mother and me. Thank you both for being my biggest fans!

Finally, I thank my wonderful husband and my best friend, Bryan Garrard, whose part in my own success is so vast that I cannot measure. I thank him for his loving support, devotion, encouragement, patience, and enormous understanding throughout this journey. Thank you, Bryan, for always being there for me and taking care of our fur babies when I am overwhelmed. You provided a sea of calm when the difficulties of graduate school were at their rockiest. Words simply cannot express.

Table of Contents

List of Figures

List of Tables

Chapter One

Introduction

Statistics, the science of making inferences regarding some population or random phenomena using data collected from representative samples, has been applied widely to disciplines such as sociology, psychology, medicine, biology, engineering, and politics. Since its beginnings around the 1700's, statistics has played an essential role in the advancement of science and society (Davidian & Louis, 2012; Stigler, 1986). Within the field of statistics, biostatistics emerges as a branch that applies and/or develops statistical methods to interpret and solve public health, biological, medical, and health sciences problems (Rosner, 2010). Through close collaborations with health care researchers, clinicians, and policy makers, biostatisticians have significant contributions in developing health policies and solving health care-related issues at both national and international levels. For example, a decades-old challenge that remains a constant debate at the political stage is to develop a health care system that has an efficient delivery and improves quality of health care for the people of United States.

Clinical outcome assessment (COA) measures provide one important aspect in assessing the quality of health care; and the development of COA measures often requires biostatisticians' expertise in statistical methods and modeling. The U. S. Department of Health and Human Services (DHHS) Food and Drug Administration (FDA) defines COAs as tools that measure specific symptoms in the patient, overall mental state, or the effects of a disease or condition on how the patient functions. Evidence collected through COA measures routinely is used by FDA to determine the treatment benefit of a drug. FDA classifies COA measures into four types: patient-reported outcome (PRO) measures (or PROMs), clinician-reported outcome (ClinRO) measures, observer-reported outcome (ObsRO) measures, and performance outcome (PerfO) measures (FDA, 2015b). Among the four types of COA measures, PROMs and ClinRO measures are used most often under the health care setting.

1.1 Clinician-Reported Outcome Measures

ClinRO measures are based on observations made by trained health care professionals on a patient's health condition; and clinical judgment or interpretation usually is involved in these potential disease- or condition-related observations (FDA, 2015a). Although the standard terminology might not be familiar to a layperson, ClinRO measures are encountered frequently in people's lives. Most of us have scheduled annual physical examinations to evaluate our overall physical wellbeing. Standard clinical procedures usually include vital signs assessments such as the measurement of temperature, pulse, blood pressure, and respiratory rate by a health care professional. When necessary, additional laboratory work such as blood test or urinalysis can be ordered to get more detailed readings on a person's health condition. All of the above-mentioned measurements are considered ClinRO measures, where the observations are made by trained health care professionals with or without the use of specific medical devices.

Despite the common use of ClinRO measures in clinical trials and/or clinical practices, there is a lack of clinical research specific to ClinRO assessments (ISPOR, 2015). The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) took initiatives to set up the Clinical Outcomes Assessment – Emerging Good Practices Task Force to address issues and make good practice recommendations for the development and evaluation of new or existing PROMs and ClinRO measures (Walton et al., 2015). The ISPOR initiatives closely align with guidelines published by FDA on COA measures. Definitions provided by the ISPOR task force classify ClinRO measures into three types: readings, ratings, and globals. Specifically, readings are presented as binary reports such as the presence of soft-tissue mass on a radiographic imaging; ratings can be either a categorical or scoring report in the form of survey

questionnaires (e.g., Positive and Negative Syndrome Scale [PANSS] for schizophrenia); and globals may involve an overall clinical judgement on the patient's health status (ISPOR, 2015).

Historically, assessing the quality of health care relied more heavily on evidence provided by ClinRO measures, whereas patient perspectives often were consulted less with the exception of gathering satisfaction feedback on care experience (National Quality Forum, 2013c). However, a comprehensive evaluation of health care involves many outcomes that expand beyond information provided by ClinRO measures. As pointed out by Cella et al. (2010), ClinRO measures (Cella et al. used the term *clinical outcome measures*; e.g., laboratory tests or radiographic imaging) have minimal immediate relevance to daily functions of patients suffering from chronic diseases (e.g., cancer) or specific symptoms that only are known to the patient (e.g., pain or fatigue). Often times patients suffering from chronic diseases may prioritize the quality of life over disease survival. In addition, an abundant body of evidence in the literature has suggested discrepancies between patient and clinician perspectives on certain disease outcomes. Some examples include the report of symptomatic toxicities such as adverse events during cancer treatment (Basch, Bennett, & Pietanza, 2011); the significance or multidimensionality of fatigue in rheumatoid arthritis (Sanderson & Kirwan, 2009); and symptoms of depression in psychotherapy research (Cuijpers, Li, Hofmann, & Andersson, 2010). Thus, patient perspectives on and beyond care experiences need to be included to assess fully the quality of health care.

1.2 Patient-Reported Outcome Measures

Patient-centered care has been recognized as a U.S. national priority for improving quality of health care (Institute of Medicine, 2001). As health care rapidly evolves into a patient-centeredness care model, the development of reliable and valid PROMs plays a critical role in

translational research and the promotion of quality care for the general population. FDA (2015a) defines PROMs as measurements based on information directly collected from the patient regarding the status of health condition without revision or interpretation by anyone else (including the clinician). In practice, PROMs often are designed as survey questionnaires with ordinal response scales, and the development of such instruments must go through rigorous testing to ensure the psychometric integrity of the instruments (Dawson, Doll, Fitzpatrick, Jenkinson, & Carr, 2010). Detailed guidelines on the development of any new or adapted PROMs have been released by several national and authoritative entities such as the National Institute of Health (NIH; Cella et al., 2010), FDA (2009), the National Quality Forum (NQF; 2013c), and the Patient-Centered Outcomes Research Institute (PCORI; 2012). Such stringent requirements are necessary as data collected using PROMs commonly are used as primary or secondary endpoints in clinical trials and studies of humans (FDA, 2014).

Over the years many PROMs have been developed and implemented in clinical trials and/or routine clinical practice, with the mission of promoting patient-centered care, supporting outcomes that patients value, and including patients in the health care-related decision-making process. Some well-established PROMs include health-related quality of life (HRQOL) questionnaires (e.g., Neuro-Qol; Gershon et al., 2012) and the Center for Epidemiologic Studies Depression Scale (CESD; Eaton, Smith, Ybarra, Muntaner, & Tien, 2004; Radloff, 1977). However, literature has suggested a lack of clarity regarding the full potential of PROMs in clinical practice (Marshall, Haywood, & Fitzpatrick, 2006), and a lack of precision and standardization among current measures (Cella et al., 2010). To address these concerns, national initiative such as the NIH Patient-Reported Outcomes Measurement Information System (PROMIS™) has been established to evaluate and develop efficient and flexible PROMs that are

publicly available (Cella et al., 2010). The PROMs developed through the PROMIS initiative include domains such as pain interference, fatigue, anxiety, and peer relationships that encompass a patient's overall wellbeing from the mental health, physical health, and social health perspectives (Gershon, Rothrock, Hanrahan, Bass, & Cella, 2010).

In addition, FDA has launched the COA qualification program formally to qualify potential COA measures for use in exploratory studies, or as primary or secondary endpoints in clinical trials. To date, the agency successfully has qualified one PROM called Exacerbations of Chronic Pulmonary Disease Tool (EXACT), developed to measure symptoms of acute bacterial exacerbation of chronic bronchitis in patients with chronic obstructive pulmonary disease (FDA, 2015b). Although qualification is not required for a PROM to be used in studies, the formal qualification process has demonstrated the agency's goal in improving outcome assessments and ensuring faster delivery of effective and safe treatments for the patients (FDA, 2014). One example of a PROM that formally was not qualified yet supported the successful approval of the drug Jakafi® is the modified Myelofibrosis Symptom Assessment Form (MFSAF) version 2.0 diary (Verstovsek et al., 2012). This novel instrument is the first developed PROM that followed FDA's guideline on the development of PRO instruments for PRO-based product labeling claim (Deisseroth et al., 2012; Zagadailov, Fine, & Shields, 2013).

1.3 Current Studies

Whether the research focus is PROMs, ClinRO measures, or COA measures in general, it is the ultimate goal for health care researchers, clinicians, and regulatory bodies to translate research findings into clinical applications and promote public awareness, thus improving the quality of life for the general population. As previously mentioned, PROMs and many ClinRO

measures are designed as psychometric instruments with ordinal response scales. However, under the current regulatory guidelines of FDA and NIH, PROMs are developed using classical (i.e., frequentist) psychometric methodologies that often are time-consuming and challenged by small samples (e.g., in cases of rare diseases). This results in substantial delays in the dissemination and transition of research findings into clinical practice. An efficient and reliable Bayesian method will offer researchers and clinicians an alternative in future PROMs development for small populations while maintaining the psychometric integrity of the instrument.

In addition, ordinal data are the most common form of data acquired from PROMs. The psychometric evaluation of PROMs, specifically the validity assessment, requires researchers to implement item response theory (IRT) models, an appropriate alternative to the classical ordinal confirmatory factor analysis (CFA). The assessment of IRT model fit is identified as both challenging and underdeveloped in the literature (Sinharay & Johnson, 2003; Sinharay, Johnson, & Stern, 2006). Therefore, it is valuable to evaluate the Bayesian alternative approach for small samples through real data applications and to investigate an appropriate method for comparing Bayesian IRT models in PROMs development.

Although the primary focus for the current studies is on the development of PROMs for small populations, the psychometric evaluation of a ClinRO measure also will be highlighted using classical psychometric approaches. One common adverse event experienced by patients in hospitals is falls. Research conducted by Shorr et al. (2008) has indicated that approximately 30% of falls result in injury, particularly among older adults. Over the years patient fall reporting has been improved remarkably through the utilization of standardized definitions; yet, injury falls reporting rarely has been examined. A NQF-endorsed falls with injury measure is assessed for its

reliability and validity to support hospitals' fall prevention efforts and future injurious falls research.

The remainder of this dissertation is organized as follows. Chapter 2 reports the publication that introduces an innovative Ordinal Bayesian Instrument Development (OBID) method for PROMs development with small samples. The performance of OBID is evaluated by applying the method to both simulated data and real data. In Chapter 3, the manuscript submitted for publication describes the OBID approach that is evaluated further with two breast cancer-related PROMs instrument development studies to assess prior selection for IRT model parameters and subject content experts' bias toward the relevancy of items. The Chapter 4 manuscript describes the psychometric assessment of a NQF-endorsed ClinRO measure—falls with injury. The dissertation concludes with summary and future studies in Chapter 5.

Chapter Two

A Novel Method for Expediting the Development of Patient-Reported Outcome Measures and an Evaluation of Its Performance via Simulation

Lili Garrard, Larry R. Price, Marjorie J. Bott, and Byron J. Gajewski

Garrard, L., Price, L. R., Bott, M. J., & Gajewski, B. J. (2015). A novel method for expediting the development of patient-reported outcome measures and an evaluation of its performance via simulation. *BMC medical research methodology, 15*(1), 77.

Abstract

Developing valid and reliable patient-reported outcome measures (PROMs) is a critical step in promoting patient-centered health care, a national priority in the U.S. Small populations or rare diseases often pose difficulties in developing PROMs using traditional methods due to small samples. To overcome the small sample size challenge while maintaining psychometric soundness, we propose an innovative Ordinal Bayesian Instrument Development (OBID) method that seamlessly integrates expert and participant data in a Bayesian item response theory (IRT) with a probit link model framework. Prior distributions obtained from expert data are imposed on the IRT model parameters and are updated with participants' data. The efficiency of OBID is evaluated by comparing its performance to classical instrument development performance using actual and simulation data. The overall performance of OBID (i.e., more reliable parameter estimates, smaller mean squared errors (MSEs) and higher predictive validity) is superior to that of classical approaches when the sample size is small (e.g. less than 100 subjects). Although OBID may exhibit larger bias, it reduces the MSEs by decreasing variances. Results also closely align with recommendations in the current literature that six subject experts will be sufficient for establishing content validity evidence. However, in the presence of highly biased experts, three experts will be adequate. This study successfully demonstrated that the OBID approach is more efficient than the classical approach when the sample size is small. OBID promises an efficient and reliable method for researchers and clinicians in future PROMs development for small populations or rare diseases.

Keywords: OBID, Bayesian psychometrics, ordinal data analysis, Bayesian IRT, patient-reported outcome measures, PROMs

2.1 Background

The Institute of Medicine (IOM; 2001) released a landmark report, *Crossing the Quality Chasm*, which highlighted patient-centered care as one of the six specific aims (the others being safety; effectiveness; timeliness; efficiency; and equity) that defined quality health care. To promote patient-centered care, national entities such as the National Institute of Health (NIH; Cella et al., 2010), the U. S. Department of Health and Human Services (DHHS) Food and Drug Administration (FDA; 2009) , the National Quality Forum (NQF; 2013c), and the Patient-Centered Outcomes Research Institute (PCORI; 2012) have published specific guidelines on the development of patient-reported outcome measures (PROMs). The guidelines unanimously emphasize the critical requirement of rigorous psychometric testing for any new or adapted PROMs that often are designed as survey instruments. PROMs serve a critical role in translational research as data collected using PROMs are commonly used as primary or surrogate endpoints for clinical trials and studies in humans, which are essential for promoting both clinical application and public awareness. However, the lengthy process of developing valid and reliable psychometric instruments (e.g., PROMs) is recognized as one of the greater barriers for disseminating and transitioning research findings into clinical practice in a timely manner.

For decades classical instrument development methodologies (e.g., frequentist approach to factor analysis that ignores prior information regarding item reliability) dominated the psychometric literature (Pett, Lackey, & Sullivan, 2003). Bayesian methods have been severely limited until modern computation techniques provided researchers the capacity to employ Bayesian inference in actual applications (Johnson & Albert, 1999). As Bayesian inference becomes more popular, limitations arise with the use of classical (i.e. frequentist) methods when developing instruments or PROMs for small populations (e.g., in cases of rare diseases). Since it

is not the intent of the authors to provide a comprehensive review of both classical and Bayesian statistical approaches, we focus our discussions on two co-existing issues with the classical approach to confirmatory factor analysis (CFA) in establishing evidence of construct validity: (a) the requirement of large samples, and (b) modeling ordinal data as continuous.

Two essential components of establishing evidence that scores acquired by an instrument exhibit score validity include content and construct-related evidence (AERA, APA, & NCME, 2014; Nunnally & Bernstein, 1994). Subject experts' opinions are typically consulted in evaluating the content of items, such as how well the items match the empirical indicators of the construct(s) of interest, and the relevancy and clarity of the items. The items evolve through rigorous revision (e.g., iteratively through pilot-testing with a small representative sample of respondents) until the instrument is deemed ready for establishing construct validity evidence through a statistical technique such as factor analysis. It is a common practice to conduct expert evaluation for content analysis; however, under the classical setting data collected from the experts are not utilized in establishing construct validity as content validity focuses on the instruments rather than measurements (Messick, 1989). The expert and participant data are analyzed separately, which results in potential loss of information and leads to the increasing demand for a large participant sample.

There is no consensus among health care researchers regarding the number of subjects required for CFA. Knapp and Brown (1995) list several competing rules regarding the number of subjects required and argue that original studies on factor analysis (e.g., Thurstone, 1947) only assumed very large samples relative to the number of items, and made no recommendations on a minimum sample size. Pett et al. (2003) make the recommendation of at least 10 to 15 subjects per item, a commonly suggested ratio in psychometric literature. However, Brown (2014) urges

researchers to not rely on these general rules of thumb and proposes more reliable model-based (e.g., Satorra-Saris's method) and Monte Carlo methods to determine the most appropriate sample size for obtaining sufficient statistical power and precision of parameter estimates. A recent systematic review study on sample size used to validate newly-developed PROMs reports that 90% of the reviewed articles had a sample size ≥ 100, whereas 7% had a sample size ≥ 1000 (Anthoine, Moret, Regnault, Sébille, & Hardouin, 2014). In addition, Weenink, Braspenning, and Wensing (2014) explore the potential development of PROMs in primary care using seven generic instruments. The authors report challenges of low response rates to questionnaires (i.e., small sample), and that a replication in larger studies would require a sample size of at least 400 patients.

Apart from the large sample issue, the other issue concerns how data are analyzed using traditional approaches. The most common form of data acquired from measurement instruments in the social, behavioral, and health sciences are ordinal; however, such data often are analyzed without regard for their ordinal nature (Johnson & Albert, 1999). The practice of treating ordinal data as continuous is considered a controversy and has generated debates in the psychometric literature (Knapp, 1990). With solid theoretical developments in ordinal data modeling, it is considered best practice to use modeling techniques that treat ordinal data as ordinal. Structure equation modeling (SEM) with categorical variables first was introduced by B. Muthén (1984) in a landmark study that revolutionized psychometric work. Although techniques for handling ordinal data in latent variable analysis have been incorporated into several commercial statistical software (e.g., Mplus) since the 1980's, it is only in 2012 that the free R package *lavaan* incorporated the weighted least squares means- and variance-adjusted (WLSMV) estimator for performing ordinal CFA during its version 0.5-9 release (R Core Team, 2015; Rosseel, 2012).

Ordinal CFA offers new insight for modeling ordinal data under the classical setting; yet it is still challenged by small samples, as we will show in this study. A more complete solution is needed to resolve both limitations and still provide reliable model estimates.

New methods proposed by Gajewski, Price, Coffland, Boyle, and Bott (2013) and Jiang et al. (2014) use Bayesian approaches to resolve the sample size limitation of traditional CFA. The Integrated Analysis of Content and Construct Validity (IACCV) approach establishes a unified model that seamlessly integrates the content and construct validity analyses (Gajewski et al., 2013). *Prior* distributions derived from content subject experts' data are updated with participants' data to obtain a *posterior* distribution. Under the IACCV approach, some of the response burden from the participants can be alleviated by using experts; thus fewer participants are needed to achieve the desired validity evidence in developing instruments. Using both simulation data and real data, Bayesian Instrument Development (BID; Jiang et al., 2014) advances the theoretical work of IACCV by demonstrating the superior performance of BID to that of classical CFA when the sample size is small. BID also advances the practical application of IACCV by incorporating the methodology into a user-friendly GUI software that is shown to be reliable and efficient in a clinical study for developing an instrument to assess symptoms in heart failure patients. Although BID has shown great potential, the method is limited by the assumption of continuous participant response data. As previously mentioned, many clinical questionnaires data are collected as ordinal or binary (a special type of ordinal data). Given this fact, there is an urgent need to adapt the BID approach for ordinal responses.

In this article, we propose an Ordinal Bayesian Instrument Development (OBID) approach within a Bayesian item response theory (IRT) framework to further advance BID methodology for ordinal data. On first glance, the current study appears to be a straightforward

extension from previous studies; however it differs from previous studies and contributes to the literature from several perspectives. First, as previously mentioned, ordinal or binary data are the most common form of data collected using clinical instruments. The underlying distribution assumption required by continuous data modeling is often violated due to skewed responses. Our study effectively promotes the proper usage of ordinal data modeling methods and brings awareness to a broader audience regarding the psychometric integrity of the measurement, which is essential for the development of PROMs and clinical trial outcomes. Although several simulation studies on Bayesian IRT models have been discussed in the literature, the studies arbitrarily select non-informative or weakly informative priors for model parameters without a clear elicitation process (e.g., Arima, 2015; Fox & Glas, 2001). Alternatively, our approach is distinct because we leverage experts in elicitation of the priors for the IRT parameters. Second, the consideration of the predictive validity of the instrument (Nunnally & Bernstein, 1994) that is often neglected in the literature is addressed here. These important steps are implemented in the simulation study for contribution to the methodological literature.

Results from our approach also have several practical implications to the development of PROMs, as OBID overcomes the small sample size (e.g., patients from small populations) challenge while maintaining psychometric integrity. Special considerations for reducing the resource and cost burden incurred by researchers and clinicians are provided through the usage of fast and reliable free R packages to implement the OBID methodology. In our approach, a Markov chain Monte Carlo (MCMC) procedure is implemented to estimate the model parameters; we provide general guidelines for selecting tuning parameters required in the MCMC procedure for achieving appropriate acceptance/rejection rates. Our proposed method demonstrates that the overall performance of OBID (i.e., more reliable parameter estimates,

smaller mean squared errors (MSE) and higher predictive validity) is superior to that of ordinal CFA when the sample size is small. Most importantly, OBID promises an efficient and reliable method for researchers and clinicians in future PROM development.

2.2 Methodology

OBID further advances the work of Jiang et al. (2014) that expands IACCV of Gajewski et al. (2013), by adapting the BID methodology for ordinal scale data. Here we demonstrate the OBID approach using a unidimensional (i.e., single factor) psychometric model and refer interested readers to Gajewski *et al.* and Jiang *et al.* for a detailed description of the general model and the BID approach. In addition, we use a similar model and incorporate mathematical notation as presented in Jiang *et al.* to maintain some level of consistency between both studies.

2.2.1 Bayesian IRT Model

Prior to introducing the OBID model, it is important to clarify that both OBID and BID are CFA-based approaches. IRT is a psychometric technique that provides a *probabilistic* framework for estimating how examinees will perform on a set of items based on their ability and characteristics of the items (Price, in press). IRT is a model-based theory of statistical estimation that conveniently places persons and items on the same metric based on the probability of response outcomes. Traditional factor analysis is based on a *deterministic* model and does not rest on a probabilistic framework. Here we provide a probabilistic connection between our approach and IRT, by using Bayesian CFA, including an inherently probabilistic framework. From a modeling perspective, IRT is the ordinal version of traditional factor analysis. When all manifest variables are ordinal, the traditional factor analysis model is equivalent to a

two-parameter IRT model with a probit link function (Johnson & Albert, 1999; Quinn, 2004). The two-parameter IRT model with the probit link can be written as

$$y_{ij} = c \text{ if } y_{ij}^{*} \in \left(T_{j(c-1),} T_{jc}\right]; \;\; i = 1, \ldots, N, \; j = 1, \; \ldots, \; P, \; c = 1, \ldots, \; C_j \tag{2.1}$$

$$y_{ij}^{*} = \alpha_j + \lambda_j f_i + \varepsilon_{ij}; \;\; f_i \sim \; N(0,1), \varepsilon_{ij} \sim N(0,1), i = 1, \ldots, \; N, \; j = 1, \; \ldots, \; P, \tag{2.2}$$

where y_{ij} is the *i*th participant's ordinal response to the *j*th item; and C_j is the total number of response categories for item *j* (e.g., a five-point Likert scale). The ordinal response y_{ij} is linked to y_{ij}^{*}, an underlying continuous latent variable that follows a normal distribution, through a set of $C_j - 1$ ordered cut-points, T_{jc}, on y_{ij}^{*}. The probability of a subject selecting a particular response category is indicated by the probability that y_{ij}^{*} falls within an interval defined by the cut-points T_{jc}. In IRT, the continuous latent variable y_{ij}^{*} is characterized by two item-specific parameters: α_j, the negative difficulty parameter for the *j*th item and λ_j, the discrimination parameter for item *j*. In addition, the underlying latent ability f_i of the subjects is constrained to follow a standard normal and ε_{ij} is the measurement error (Johnson & Albert, 1999).

To see the equivalence between the IRT model and traditional factor analysis model, note that a classical unidimensional factor analysis model can be expressed as

$$z_{ij}^{*} = \rho_j f_i + e_{ij}; \;\; i = 1, \ldots, N, \; j = 1, \ldots, P, \tag{2.3}$$

where z_{ij}^{*} represents the standardized y_{ij}^{*} from equations 2.1 and 2.2; f_i is the *i*th participant's factor score for the domain; ρ_j is the factor loading or item-to-domain correlation for the *j*th item; and e_{ij} represents the measurement errors or sometimes referred to as latent unique factors or residuals. f_i is assumed to follow a standard normal distribution, which implies that

$e_{ij} \sim N(0,\ 1-\rho_j^2)$ where ρ_j^2 is the reliability of the *j*th item. The standardization of y_{ij}^* is expressed by

$$\frac{y_{ij}^*-\alpha_j}{\sqrt{1+\lambda_j^2}} = \frac{\lambda_j}{\sqrt{1+\lambda_j^2}} f_i + \frac{\varepsilon_{ij}}{\sqrt{1+\lambda_j^2}};\quad i = 1, \dots, N,\ j = 1, \dots, P, \tag{2.4}$$

such that the IRT model parameter λ_j can be interpreted interchangeably through the item-to-domain correlations ρ_j using the following expressions

$$\lambda_j = \frac{\rho_j}{\sqrt{1-\rho_j^2}} \tag{2.5}$$

$$\rho_j = \frac{\lambda_j}{\sqrt{1+\lambda_j^2}}\,. \tag{2.6}$$

Equations 2.5 and 2.6 can be interpreted such that an item that well-discriminates among individuals with different abilities also will have a high item-to-domain correlation. The true Bayesian application comes from specifying appropriate prior distributions on the IRT parameters, which leads us into the essence of the OBID method.

2.2.2 OBID – Expert Data and Model

Eliciting subject experts' perception regarding the relevancy of each item to the domain (construct) of interest is a common practice to aid in verifying content validity evidence. For example, during instrument development, a logical structure is developed and applied in a way that maps the items on the test to a content domain (AERA, APA, & NCME, 2014). In this way, the relevance of each item and the adequacy with which the set of items represents the content domain is established. To illustrate, a panel of subject experts are asked to review a set of potential items and instructed to provide response for questions such as "please rate the relevancy of each item to the overall topic of [domain]." The response options are generally

designed on a four-point Likert scale that ranges from “not relevant” to “highly relevant.” Gajewski et al. (2012) laid important groundwork from an empirical perspective by demonstrating the approximate equivalency of measuring content validity using relevance scales versus using correlation scales. In other words, content validity oriented evidence can be statistically interpreted as a representation of the experts’ perceptions regarding the item-to-domain latent correlation (Jiang et al., 2014).

Continuing the notations from Jiang *et al.*, suppose the expert data are collected from a panel of $k = 1, \dots, K$ experts that respond to $j = 1, \dots, P$ items. Let $\boldsymbol{X}$ denote the $K \times P$ matrix of observed ordinal responses where the x_{jk}th entry represents the kth expert’s opinion regarding the relevancy of the jth item to its assigned domain. Similarly, the kth expert’s latent correlation between the jth item and its respective domain is denoted by ρ_{jk} and is related to x_{jk} using the following function, with correlation cut-points suggested by Cohen (1988):

$$x_{jk} = \begin{cases} 1 \text{ "not relevant"} & \text{if } 0.00 \leq \rho_{jk} < 0.10 \\ 2 \text{ "somewhat relevant"} & \text{if } 0.10 \leq \rho_{jk} < 0.30 \\ 3 \text{ "quite relevant"} & \text{if } 0.30 \leq \rho_{jk} < 0.50 \\ 4 \text{ "highly relevant"} & \text{if } 0.50 \leq \rho_{jk} \leq 1.00 \end{cases}. \tag{2.7}$$

A sensitivity analysis conducted by Gajewski et al. (2012) demonstrated the approximate equivalency of using correlation scale and using relevancy scale to measure content validity, under both equally-spaced (i.e., $0.00 \leq \rho_{jk} < 0.25$, $0.25 \leq \rho_{jk} < 0.50$, $0.50 \leq \rho_{jk} < 0.75$, and $0.75 \leq \rho_{jk} < 1.00$) and unequally spaced (i.e., equation 2.7) cut-points assumptions. One of the reviewers pointed out that under certain circumstances, the equally-spaced transformation might be more appropriate (e.g., a panel with moderate level of expertise in the area of interest) (Gajewski et al., 2012). However, the results were based on unexpected secondary findings, which require further confirmation in a more thorough study (Gajewski et al., 2012). For the

purpose of the current study, we want to primarily focus on showcasing a proper method of establishing evidence for construct validity using carefully selected "true" subject experts. For developing PROMs, the level of expertise of the selected subject experts' has a direct impact on the validity of the measurement instrument.

In our assumed single factor model, the item-to-domain correlation based on pooled information from all experts can be denoted by $\rho_j = corr(f, z_j)$, where f represents the domain factor score and is typically assumed to follow a standard normal distribution; and z_j represents the standardized response of item *j*. To ensure the proper range of correlations, Fisher's transformation is used to transform ρ_j and we denote μ_j as

$$\mu_j = g(\rho_j) = \frac{1}{2} log \frac{1+\rho_j}{1-\rho_j}. \quad (2.8)$$

A hierarchical model that combines all experts and includes all items is defined by

$$g(\rho_{jk}) = g(\rho_j) + e_{jk}, \quad (2.9)$$

where $e_{jk} \sim N(0, \sigma^2)$. Following the BID model, the prior distribution of the experts after Fisher's transformation is approximately normal and can be expressed by

$$\mu_j = g(\rho_j) \sim N\left(g(\rho_{0j}), \frac{1}{n_{0j}}\right), \quad (2.10)$$

where $g(\rho_{0j})$ is the transformed prior mean item-to-domain correlation; and $n_{0j} = 5 \times K$ is the prior samples size such that each expert is equivalent to approximately five participants (Jiang et al., 2014). This approximation is based on a weighted average from previous study findings by Gajewski et al. (2012), Gajewski et al. (2013), and Jiang et al. (2014). The prior sample size n_{0j} can be approximated by computing the ratio of the variance of the subject experts' transformed ρ_j and the variance of the participants' transformed ρ_j (i.e., using a flat prior). The "five participants" assumption will be further evaluated as more data become available. Moreover, the

current approximation is solely needed to help execute the simulation study and not used within any real data application.

Informative priors only should be used when appropriate content information is available. When items are substantially revised without further review from subject experts, flat priors should be used. Although eliciting prior distribution from subject experts is highlighted, we are not restricted solely to this approach. When reliable and relevant external data are available (i.e., not necessarily experts), a different data driven approach can be utilized. For instance, developing PROMs for pediatric populations can be challenging due to low disease incidence in children, thus resulting in small samples. Reliable evidence from the adult populations can be treated as a "general prior" for establishing construct validity in the pediatric populations.

2.2.3 OBID – Participant Data and Model

Establishing evidence of score validity involves integrating various strategies or techniques culminating in a comprehensive account for the degree to which existing evidence and theory support the intended interpretation of scores acquired from the instrument (Price, in press). From a purely psychometric or statistical perspective, establishing content validity evidence has traditionally been carried out separately from establishing evidence of construct validity. Importantly, the OBID approach more closely aligns with current practice forwarded by the American Educational Research Association (AERA), American Psychological Association (APA) and the National Council on Measurement in Education (NCME; AERA, APA, & NCME, 2014) regarding an integrated approach to establishing evidence for score validity in relation to practical use. OBID seamlessly integrates content and construct validity analyses into a single process, which alleviates the need for a large participant sample. The previously introduced IRT

with a probit link model, expressed by equations 2.1 and 2.2, is used to model the ordinal participant responses. The likelihood for y_{ij}^* is

$$L(\boldsymbol{y}^*|\boldsymbol{\alpha},\boldsymbol{\lambda},\boldsymbol{f}) = \prod_{i=1}^{N}\prod_{j=1}^{P} N(y_{ij}^*|\alpha_j + \lambda_j f_i\ ,\ 1). \tag{2.11}$$

By equations 2.5, 2.8, 2.10 and the delta method, we specify the prior distribution of the item discrimination parameter λ_j through a normal approximation where

$$\lambda_j \sim N\left(\frac{\exp(2\mu_j)-1}{2\exp(\mu_j)}, \frac{\{\exp(2\mu_j)+1\}^2}{4n_{0j}\exp(2\mu_j)}\right). \tag{2.12}$$

Since the item-to-domain correlation ρ_j does not depend on the negative item difficulty parameter α_j, we assign the prior $\alpha_j \sim N(0,1)$ according to recommendations made by Johnson and Albert (1999). The full posterior distribution is

$$\pi(\boldsymbol{\alpha},\boldsymbol{\lambda}|\boldsymbol{y}^*,\boldsymbol{f}) = \prod_{i=1}^{N}\prod_{j=1}^{P} N(y_{ij}^*|\alpha_j + \lambda_j f_i\ ,\ 1) \times \prod_{i=1}^{N} N(f_i|0,1) \times \prod_{j=1}^{P} N\big(\alpha_j|0,1\big)$$

$$\times \prod_{j=1}^{P} N\left(\lambda_j \left| \frac{\exp(2\mu_j)-1}{2\exp(\mu_j)}, \frac{\{\exp(2\mu_j)+1\}^2}{4n_{0j}\exp(2\mu_j)}\right.\right) \times \prod_{j=1}^{P} N\left(\mu_j \left| \mu_{0j},\ \frac{1}{n_{0j}}\right.\right). \tag{2.13}$$

2.2.4 OBID Model Estimation

The integration of content and construct validity analyses requires us to calculate the posterior distribution of the expert data and use the posterior inferences as priors for the participant model parameters, as expressed in equation 2.13. Prior to eliciting expert opinions, it is natural to assume that no information exists regarding the items. Thus, flat or non-informative priors can be specified in equations 2.9 and 2.10 such that $\sigma^2 \sim IG(0.00001, 0.00001)$ and $\mu_j = g(\rho_j) \sim N(0,3)$. The MCMC procedure is implemented in the free software WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) to estimate the posterior distribution of λ_j based on μ_j from the experts' data. Three chains are used with a burn-in sample of 2,000 draws. The next

10,000 iterations are used to calculate the posterior inferences that form the priors of λ_j in the participant IRT model.

The estimation of λ_j's in the participant model can be obtained by using the *MCMCordfactanal* function included in the free R package *MCMCpack* (Martin, Quinn, & Park, 2011). To be specific, the R function utilizes a Metropolis-Hastings within Gibbs sampling algorithm proposed by Cowles (Cowles, 1996). Similarly, the posterior estimation of λ_j's is based on 10,000 iterations after 2,000 burn-in draws. The item-to-domain correlations ρ_j's can be subsequently calculated from the estimated λ_j's via equation 2.6. An important consideration in any MCMC procedure is the choice of a tuning parameter that influences the appropriate acceptance or rejection rate for each model parameter. According to Gelman, Carlin, Stern, and Rubin (2004) and Quinn (2004), the proportion of accepted candidate values should fall between 20- 50%. There is no standard "formula" for selecting the most appropriate tuning parameter. As Quinn suggested, users typically adjust the value of the turning parameter through trial and error. In the upcoming discussion of the simulation study, we have found that the following tuning parameter values 1.00, 0.70, 0.50, and 0.30 appear to work well for sample sizes 50, 100, 200, and 500, respectively.

2.2.5 Predictive Validity

An essential yet often neglected instrument evaluation step is the assessment of predictive validity. Predictive validity is sometimes referred to as criterion-related validity whereas the criterion is external to the current predictor instrument. From a statistical standpoint, assuming the availability of an appropriate criterion, the predictive validity is directly indicated by the size of the correlation between predictor scores and criterion scores. However, demonstrating

construct validity of an instrument may not always support the establishment of predictive validity due to factors such as range restriction, where the relevant differences on the predictor or criterion are eliminated or minimized. Thus, the performance of predictive validity depends entirely on the extent to which predictor scores correlate with criterion scores intended to be predicted (Nunnally & Bernstein, 1994; Price, in press).

In this article we compare the OBID predictive validity with that of the traditional approach. Using the test scores or the underlying latent ability parameter f_i of the subjects, the validity coefficient is defined as

$$\gamma = corr\{E(\boldsymbol{f}), \boldsymbol{f}^T\}, \tag{2.14}$$

where $E(\boldsymbol{f})$ is the posterior mean of the test scores and $\boldsymbol{f}^T$ represents the set of true test scores. In our simulation study, the criterion is assumed to be perfectly measured; thus the correlation of the test score f_i (i.e., the ability parameter) and the criterion score is the same as the validity coefficient corrected for attenuation in the criterion only.

2.3 Results

2.3.1 Simulation Study

In this section, we use simulated data to test the OBID approach by comparing its overall performance to classical instrument development, specifically through the comparison of parameter estimates, MSEs, and predictive validity. Two important assumptions are made by Jiang et al. (2014) for BID that also apply to the OBID simulation setting. First, all experts are assumed to agree in regards to interpreting the concept of correlation in their opinions about the items' relevancy; and second, the experts' data are assumed to be correlated with the participants' data with the indication of having either the same opinions or very similar opinions. In addition,

the BID study makes the assumption that the true item-to-domain correlation is $\rho^T = 0.50$ for all items. Upon careful consideration, we have decided against this assumption for the current study as in reality it is rare for all items to have the same moderate item-to-domain correlation. Thus, we employ a mixture of low, moderate, and high (i.e., 0.30, 0.50, and 0.70) true item-to-domain correlations in this simulation study. The simulation is conducted in R software version 3.1.2 (R Core Team, 2015), including additional inferences and simulation plots. OBID parameter estimation is obtained using the previously introduced *MCMCordfactanal* function in the R package *MCMCpack* (Martin et al., 2011). In addition, for comparison purposes ordinal CFA is performed using the *cfa* function in the R package *lavaan* version 0.5-17 (Rosseel, 2012).

Working with the assumed unidimensional model, a five-way factorial design is used to simulate the data. The simulation factors include number of items on the instrument (4, 6, 9) and number of response categories per item (2, 5, 7). For simplicity and demonstration purposes, we assume that all items have the same number of response categories in the current simulation. However, it is possible for items to have different number of response categories on a questionnaire. In addition, we examine the effect of expert bias using different number of participants (50, 100, 200, 500), number of subject experts (2, 3, 6, 16), and types of expert bias (unbiased, moderately biased, highly biased). We define unbiased experts as $\rho_0 = \rho^T$, moderately biased experts as $\rho_0 = \rho^T + 0.1$, and highly biased experts as $\rho_0 = \rho^T + \frac{1-\rho^T}{2}$. This design results in 432 different combinations of factors. The detailed simulation strategy is as follows:

1. Simulate standardized participant responses z_{ij}^* and convert to y_{ij}^* based on the classical factor model (equation 2.3). The true item-to-domain correlation ρ^T is specified as $\rho^T = (0.50, 0.30, 0.70, 0.50)$ for all four item scenarios, $\rho^T =$

$(0.30, 0.50, 0.70, 0.70, 0.30, 0.50)$ for all six item scenarios, and $\rho^T = (0.30, 0.50, 0.70, 0.70, 0.30, 0.50, 0.70, 0.50, 0.30)$ for all nine item scenarios.

2. Convert y_{ij}^* to ordinal responses y_{ij} using equation 2.1 and percentile-based cut points. When the number of categories is binary, or $C = 2$, the single cut point is the 50th percentile of the standard normal. When the number of categories is polytomous, or $C > 2$, the cut points are defined as the $\left(\frac{1}{C}, \ \ldots, \ \frac{C-1}{C}\right)$th percentile of the standard normal.
3. Define prior for the participant IRT model (equation 2.2) item discrimination parameter λ_j using equations 2.8, 2.10, and 2.11. Recall that we previously specify the prior for the negative item difficulty parameter α_j as $\alpha_j \sim N(0, 1)$.
4. Select appropriate tuning parameters to ensure 20-50% acceptance rate. As previously mentioned, we have found through trial and error that the following tuning parameter values 1.00, 0.70, 0.50, and 0.30 appear to work well for sample sizes *N* = 50, 100, 200, and 500, respectively.
5. Fit the IRT model on the simulated datasets created in steps 1-2 via *MCMCpack* and obtain estimates for λ_j and ρ_j using equations 2.5 and 2.6.
6. Fit the ordinal CFA model on the same simulated datasets created in steps 1-2 via *lavaan* and estimate ρ_j.
7. Perform 100 simulations for each of the scenarios defined by the simulation factors.

The simulation process for one type of expert bias takes about two days to run on an Intel Core i7 3.40 GHz computer with 32GB of RAM. In order to compare the overall performances of OBID and CFA, we calculate the average MSE of the item-to-domain correlation estimates

and the MSE of the validity coefficient estimates across 100 simulations with 5,000 MCMC iterations and 2,000 burn-in draws. We denote $\hat{\rho}_j(s)$ as the OBID posterior mean or CFA parameter estimate of the *s*th iteration and $\bar{\rho}_j = \frac{\sum_{s=1}^{100} \hat{\rho}_j(s)}{100}$. Then $MSE(\hat{\rho}_j) = \frac{\sum_{s=1}^{100}\{\hat{\rho}_j(s) - \rho_j^T\}^2}{100}$ and $\overline{MSE} = \frac{\sum_{j=1}^{P} MSE(\hat{\rho}_j)}{P}$; $\{Bias(\hat{\rho}_j, \rho_j^T)\}^2 = (\bar{\rho}_j - \rho_j^T)^2$ and $\overline{Bias^2} = \frac{\sum_{j=1}^{P}\{Bias(\hat{\rho}_j, \rho_j^T)\}^2}{P}$. For evaluating the predictive validity, we denote $\hat{\gamma}(s) = corr[E\{\hat{f}_i(s)\}, f_i^T(s)]$ as the correlation between the posterior mean of estimated factor scores and true factor scores for the *s*th iteration. As previously mentioned, we assume that the true criterion is perfectly measured such that $\gamma^T = 1$. Then $MSE(\gamma) = \frac{\sum_{s=1}^{100}\{\hat{\gamma}(s) - \gamma^T\}^2}{100}$. In addition, due to concerns about the performance of CFA with small samples, we record the frequency that ordinal CFA fails to converge and/or produces "bad" estimates such that $\rho_j \notin [-1, 1]$.

Figure 2.1 shows the average MSE of item-to-domain correlation ρ for unbiased experts when the number of items (*P*) is six. The participant sample sizes are *N* = 50, 100, 200, and 500. The numbers of response categories are *C* = 2, 5, and 7, and the numbers of experts are *K* = 2, 3, 6, and 16. The MSE for CFA does not change with the number of experts (dashed line) as the expert content validity information is not utilized under the traditional approach. Thus the prior information has no effect on the CFA estimates across different choices for the number of experts. The OBID MSE (solid line) is consistently smaller than the CFA MSE, regardless of sample size and number of response categories, demonstrating the superior performance of the OBID approach. OBID is most promising for smaller samples (e.g., *N* =50 or 100). In addition, the OBID MSE decreases as the number of experts increases, with the largest reduction occurring approximately between 3-6 experts. When the number of response categories is binary

(C = 2), we observe the largest vertical distance between the OBID MSE and the CFA MSE.

This vertical distance reduces as the number of response categories increase, due to an increase

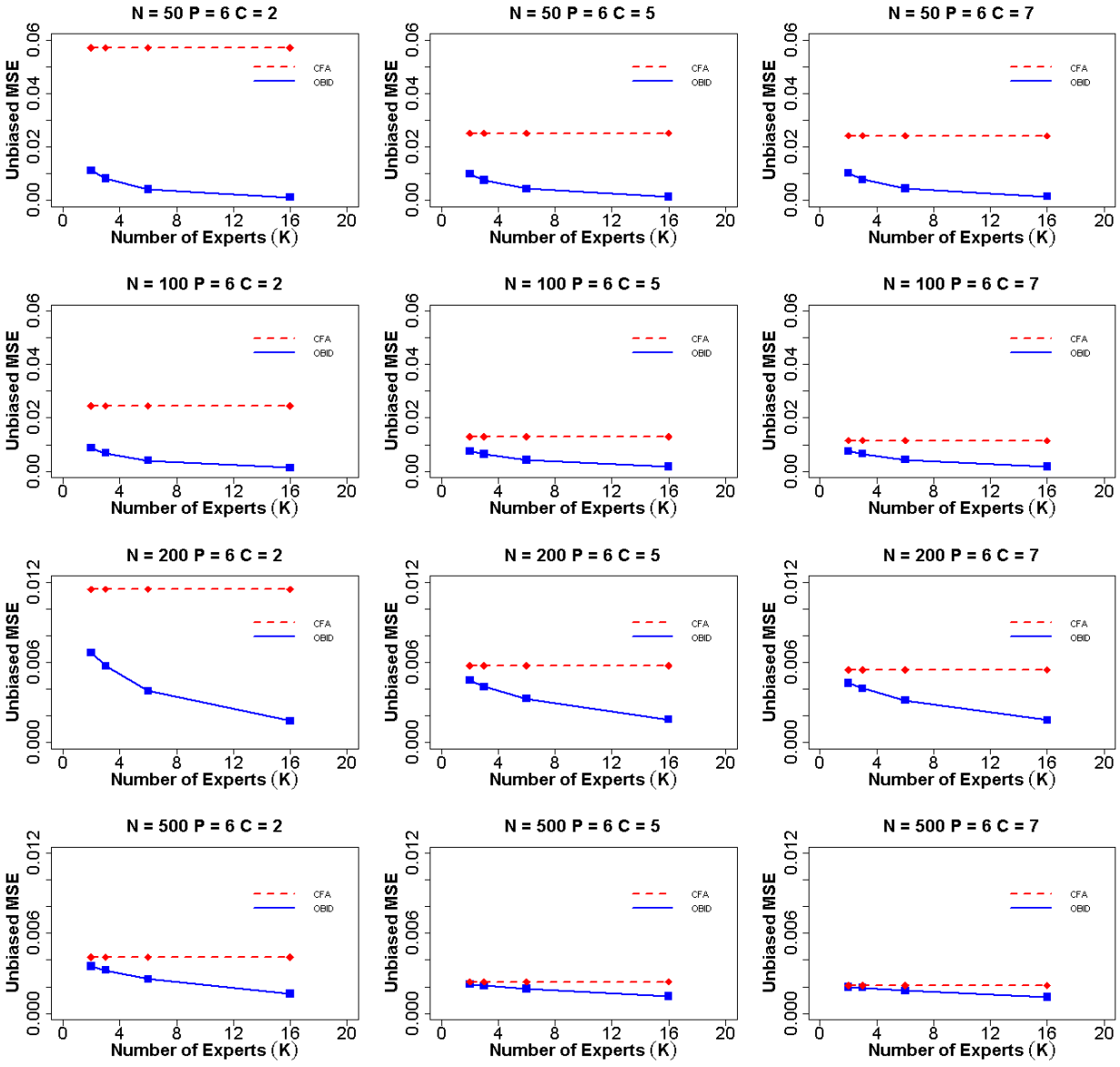


Figure 2.1. Average MSE of item-to-domain correlation ρ for six items and unbiased experts. Average mean squared error (MSE) for item-to-domain correlation ρ using OBID (solid blue line) and ordinal CFA (dashed red line) when P = 6 (number of items) and experts are unbiased $\rho_0 = (0.30, 0.50, 0.70, 0.70, 0.30, 0.50)$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.
Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

in scale information. Similarly, the MSEs for both OBID and CFA decrease as the number of response categories increase; however, the MSE graphs for the five- and seven-point scales become very similar to each other across all sample sizes. It's also expected that the MSEs for both approaches decrease as sample size increases, as a result of decreasing measurement errors. The asymptotic behavior of OBID is evaluated with sample size 500. As we expect, the two approaches produce almost identical MSEs with OBID being slightly smaller.

When experts are moderately biased (Figure 2.2), a similar overall trend is observed as that of the unbiased case. OBID continues to outperform CFA in all scenarios; however, the differences in MSEs between OBID and CFA become smaller in the moderately biased case, indicating the effect of biased priors. Additionally, the efficiency gain of the OBID approach experiences a steady increase from 2-6 experts, and gradually levels off from 6-16 experts. This indicates that with moderately biased priors, having more than six experts does not contribute to any additional gain in the efficiency of OBID. When priors are highly biased (Figure 2.3), our results support similar findings of BID (Jiang et al., 2014) where the relative efficiency of OBID compared with CFA is a function of the number of experts. In the case of a binary response option and sample size 50, OBID produces smaller MSEs than CFA, despite of the receding efficiency as the number of experts increases. OBID is most efficient with smaller samples (e.g., $N \leq 100$) and the number of experts is two or three. As number of experts increases, the impact of highly biased priors is substantial with smaller samples. The differences in MSEs between the OBID and CFA approaches exhibit similar patterns when the number of items is four or nine. MSE plots for additional simulation scenarios are included in Figures S2.1-S2.6 of the appendix.

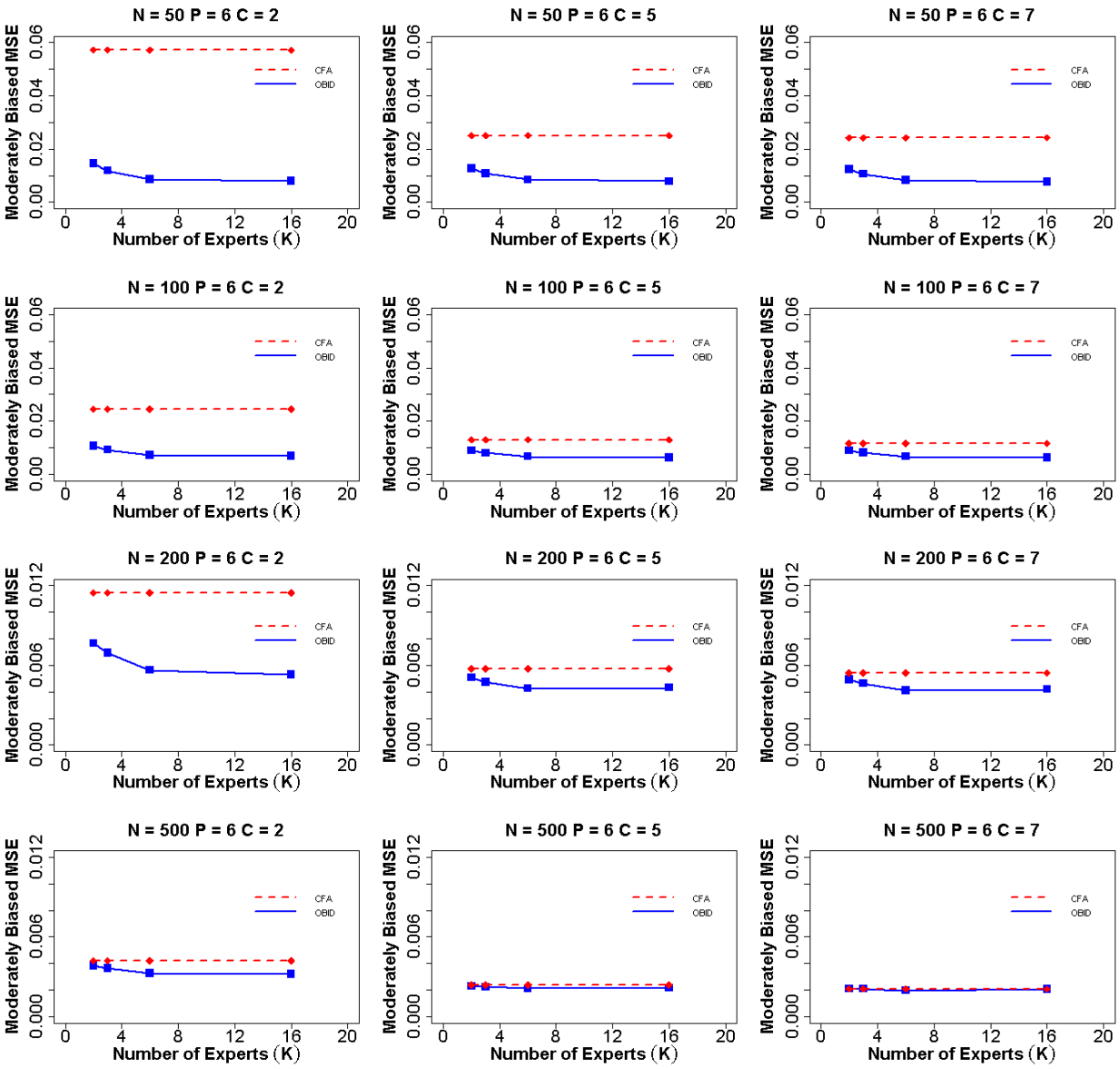


Figure 2.2. Average MSE of item-to-domain correlation ρ for six items and moderately biased experts. Average mean squared error (MSE) for item-to-domain correlation ρ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 6$ (number of items) and experts are moderately biased $\rho_0 = (0.40, 0.60, 0.80, 0.80, 0.40, 0.60)$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.
Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

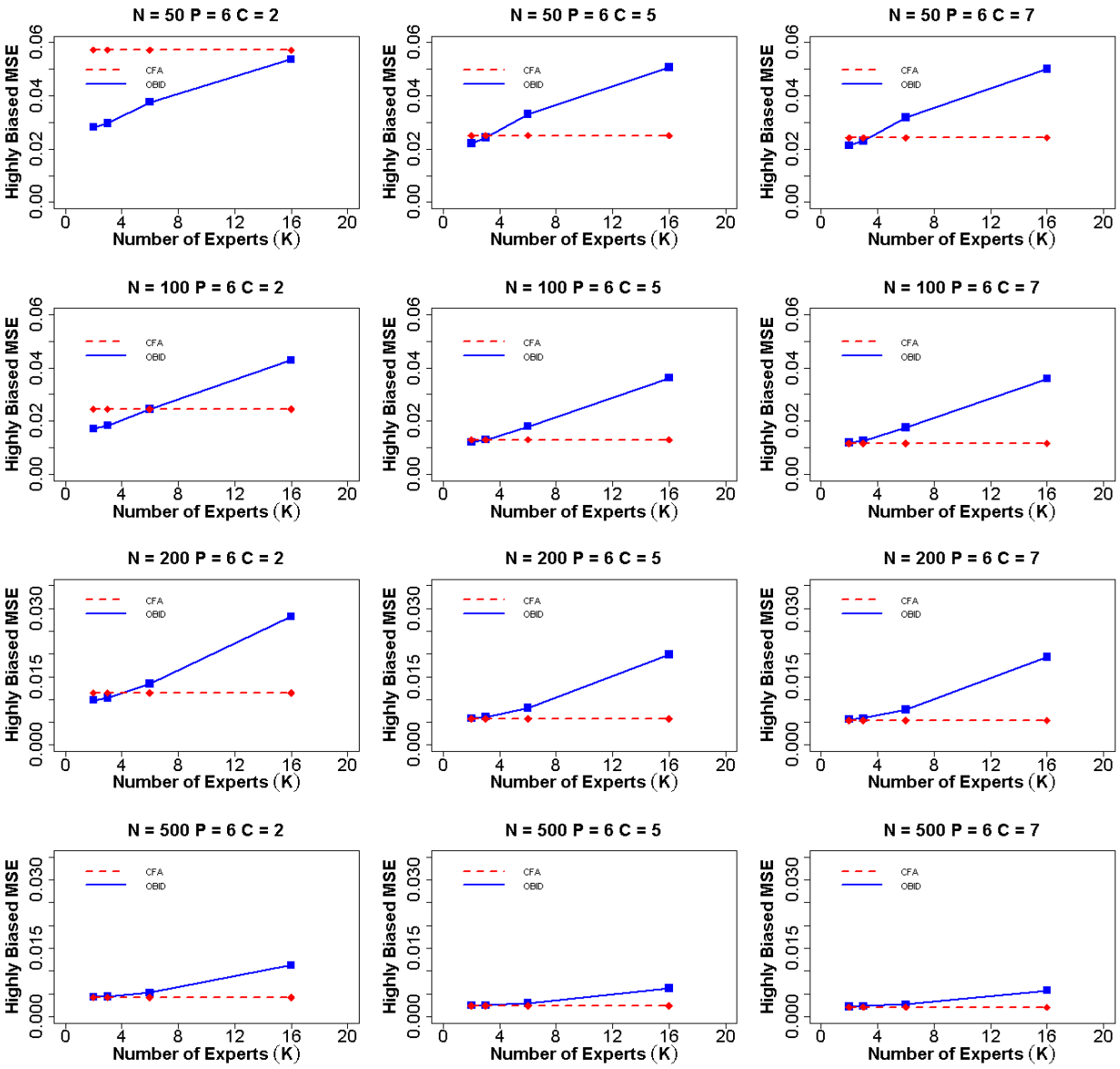


Figure 2.3. Average MSE of item-to-domain correlation ρ for six items and highly biased experts. Average mean squared error (MSE) for item-to-domain correlation ρ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 6$ (number of items) and experts are highly biased $\rho_0 = (0.65, 0.75, 0.85, 0.85, 0.65, 0.75)$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.
Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

From simply observing the graphs, one may think that although OBID is more efficient, the performance of ordinal CFA is comparable and not a bad choice. However, a close

examination of the frequency that ordinal CFA failed to converge and/or produced "bad" estimates (i.e., $\rho_j \notin [-1, 1]$) reveals limitations of the classical method with small samples. In the six item simulation example, when $N = 50$ and $C = 2$, ordinal CFA fails to converge for 2% of simulation iterations and produces out of bound correlation estimates for 21% of simulation iterations. When both sample size and number of response categories increase, although all simulation iterations converge, CFA continues to produce 1-3% out of bound correlation estimates. The four item scenarios face more challenges with convergence and reliable estimates with smaller samples. When the number of items is nine, the performance of CFA becomes more stable with only 6% out of bound estimates in the sample size 50 and binary response option case. The complete table that summarizes CFA performance can be found in Table S2.1 of the appendix. In contrast, the OBID approach consistently produces appropriate and reliable correlation estimates without any challenges using all sample sizes and response options.

Lastly we assessed the predictive validity of the two approaches under simulation settings. Under the previously mentioned assumption, the criterion is perfectly measured (i.e., the ideal target); thus the correlation of test scores f_i (i.e., the ability parameter) and criterion scores is the same as the validity coefficient corrected for attenuation in the criterion only. Figure 2.4 displays the MSEs of the validity coefficient γ computed using both OBID and CFA approaches when experts are highly biased and the number of items is six. Based on findings from Gajewski et al. (2013), the subject experts tend to overestimate the relevancy of items, resulting in highly biased item-to-domain correlations. The predictive validity of OBID is examined in the extreme case of highly biased priors with a small sample size. For 50 participants, we can clearly observe that the MSE of OBID is the smallest with a binary response option (C=2), compared with the CFA MSE. As number of response categories increases, OBID continues to have smaller MSE than that of

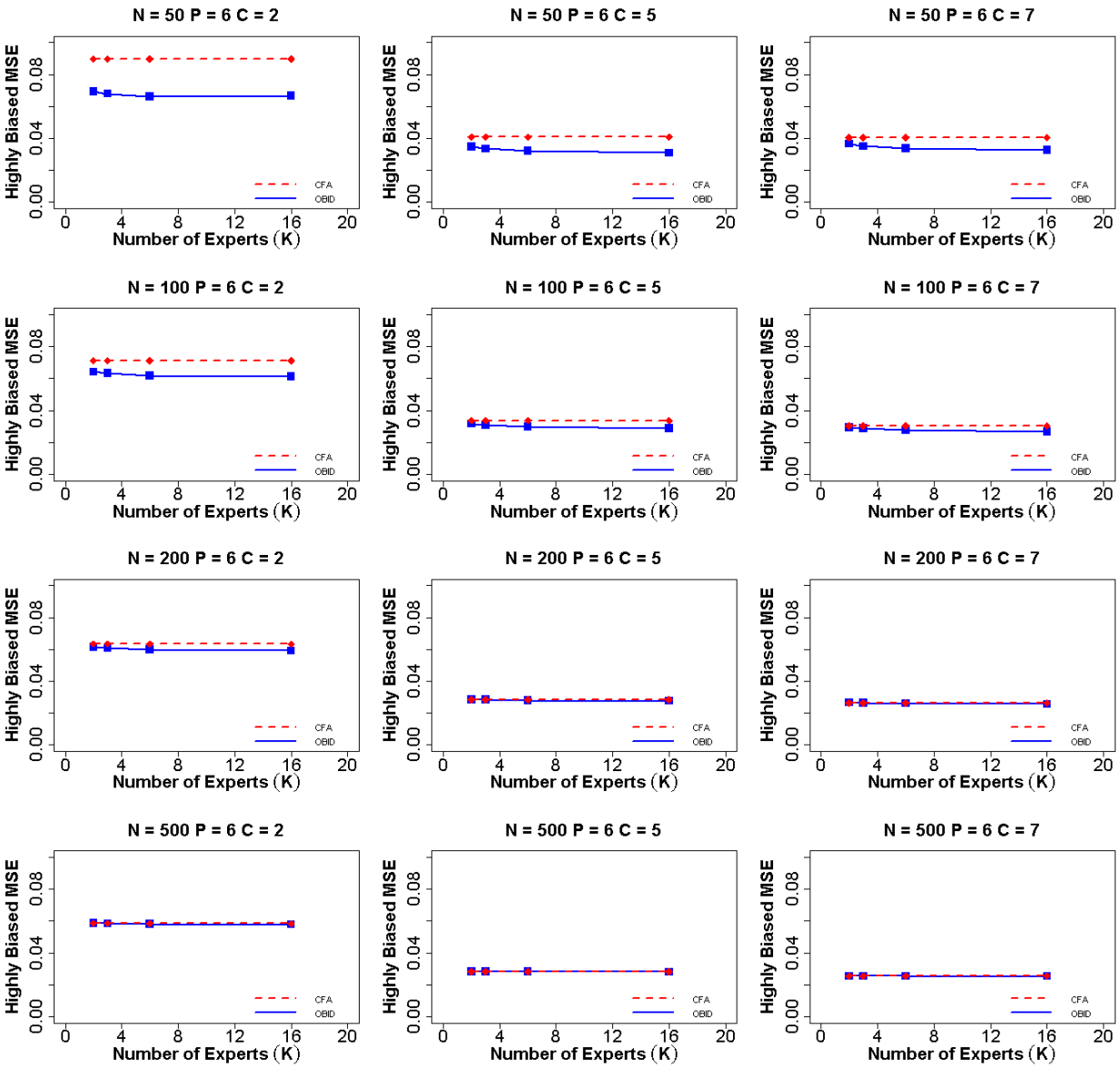


Figure 2.4. Average MSE of validity coefficient γ for six items and highly biased experts. Mean squared error (MSE) for validity coefficient γ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 6$ (number of items) and experts are highly biased $\rho_0 = (0.65, 0.75, 0.85, 0.85, 0.65, 0.75)$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.
Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

CFA, although the differences become much smaller and almost negligible. When we increase the sample size, the two approaches become almost identical in terms of MSEs. A similar trend

is observed in the four item and nine item scenarios, with corresponding plots included in Figures S2.7-S2.14 of the appendix. Prior to the simulation, we hypothesize that $MSE(\gamma_{OBID}) < MSE(\gamma_{CFA})$, $\boldsymbol{f}_{OBID}$ is more correlated with $\boldsymbol{f}^T$ than $\boldsymbol{f}_{CFA}$. The simulation results support this original hypothesis. Thus, we make the conclusion that OBID produces higher predictive validity than that of the traditional approach, especially for small samples.

2.3.2 Application to PAMS Short Form Satisfaction Survey Data

Due to scarcely available mammography-specific satisfaction assessments, researchers at a Midwestern academic medical center developed the patient assessment of mammography services (PAMS) satisfaction survey (four-factor with 20 items) and PAMS-Short Form (single factor with seven items) (Engelman et al., in review). In this section, we apply the OBID approach to complete data collected from the PAMS-Short Form instrument that was administered to 2,865 women: Hispanic (36, 1.26%), Non-Hispanic white (2,768, 96.61%), African American (34, 1.19%), and other (27, 0.94%). Participants rated their satisfaction with each of the seven items using a five-point Likert-type scale, ranging from "poor" to "excellent." In addition, six subject experts were consulted and instructed to evaluate each of the seven items on a four-point relevancy scale. The University of Kansas Medical Center's Internal Review Board (IRB) has determined that our study does not require oversight by the Human Subjects Committee (HSC), as data were collected for prior studies and they are provided to us in a de-identified fashion.

Based on the sample size for each racial/ethnic group, establishing construct validity evidence for scores for Non-Hispanic white participants is clearly adequate and traditional CFA will suffice based on the large sample. Yet, researchers are interested in establishing score-based

construct validity evidence for groups such as Hispanic/African Americans which are typically small. Classical CFA is ill-suited for such small samples; thus we apply the OBID approach for the analyses of Hispanic/African American populations. For comparison purposes, we perform OBID with experts' opinions (informative) and OBID without experts' opinions (non-informative) due to estimation challenges with traditional CFA. Flat priors are assigned for the IRT model parameters in the OBID posterior non-informative cases, in which, $\alpha_j \sim N(0,1)$ and $\lambda_j \sim N(0,4)$. In addition, based on trial and error we set the tuning parameter value required for *MCMCpack* to 2.00 for both small populations. The estimated item-to-domain correlation ρ_j and its corresponding standard error are reported in Table S2.2 of the appendix.

The non-informative OBID tends to overestimate ρ_j compared with the experts' estimated correlations (.381-.673), for both Hispanic (.570-.920) and African American (.774-.942) populations. By integrating the experts' opinions with participants' data, informative OBID produces more reliable results (Hispanic: .466-.717; African American: .495-.725) by appropriately lowering the estimated ρ_j. Although not reported, the factor score or latent variable score for each participant (i.e., individual mammography satisfaction) also is estimated. Since the factor scores are adjusted or corrected for measurement error, patients can be more accurately classified into diagnostic groups based on factor scores, and then treated as covariates in subsequent analyses. The non-informative OBID estimates tend to have slightly smaller standard errors, which can be viewed as a trade-off between the overestimated reliability ρ_j^2 and the variance. Overall, as we expect, OBID successfully produces reliable item-to-domain correlation estimates and overcomes the small sample size challenge that often causes classical CFA to fail.

2.4 Discussion

As health care moves rapidly toward a patient-centeredness care model, the development of reliable and valid PROMs is recognized as an essential step in promoting quality care. Despite of increasing public awareness, the development of PROMs using traditional psychometric methodologies often is lengthy and constrained by the large sample size requirement, resulting in substantially increased costs and resources. In this study, an innovative OBID approach within a Bayesian IRT framework is proposed to overcome both small sample size (e.g., patients from small populations or rare diseases) and ordinal data modeling limitations. OBID seamlessly and efficiently utilizes subject experts' opinions (content validity) to form the prior distributions for the IRT parameters in construct validity analysis, as opposed to using arbitrarily selected priors in other Bayesian IRT simulation studies mentioned in the introduction.

A thorough comparison between OBID and traditional CFA is provided through assessing item-to-domain correlation estimates, MSEs, and predictive validity under a simulation setting with three different types of expert bias. Simulation results across all three types of expert bias clearly demonstrate that the overall performance of OBID is most superior to that of traditional CFA when the sample size is small (i.e., ≤ 100 participants) and the instrument response option is binary. When subject experts are biased, the gain in efficiency gradually recedes for OBID as number of experts increases; and traditional CFA eventually becomes more efficient. Although not discussed in the article, the average squared bias for the item-to-domain correlation estimate also is examined across different expert biases. The corresponding plots are included in Figures S2.15-S2.23 of the appendix. A trade-off situation is observed as OBID may exhibit larger bias; yet it reduces the MSEs by decreasing variances. In addition, OBID produces higher predictive validity than that of the traditional method when the sample size is small. The

simulation results are supported by the PAMS-Short Form example where OBID is successfully applied to small Hispanic and African American populations. The de-identified PAMS-Short Form data are available in a de-identified fashion to researchers upon request through e-mail to the corresponding author of this paper. Overall, while traditional methods are restricted by small samples, OBID proves to be an efficient and reliable approach.

One limitation of this study is associated with the source of experts' information used in the PAMS-Short Form example. Opinions from the six content experts were originally consulted with the purpose of validating the PAMS instrument for the American Indian women population. Although the same set of survey items was administered to all American Indian, Hispanic, and African American populations, potential bias could be introduced due to the original focus of content experts. Nonetheless, as previously mentioned, reliable information collected from the six experts can still be utilized to form a "general prior" in establishing construct validity for Hispanic and African American populations. Another limitation of the study comes from the elicitation of content validity using relevance scales. Although Gajewski et al. (2012) has demonstrated the appropriateness of measuring content validity using relevance scales, the equivalency with measuring content validity using correlation scales is approximate, which may have an effect on the parameter estimation. A third limitation of the study comes from the approximate normal distribution assumption that we made regarding the prior distribution of the experts after Fisher's transformation. As pointed out by one of the reviewers, potential disagreements among selected subject experts may occur, which can cause the expert opinion to follow a bimodal (i.e., two groups of experts with opposite views) or even trimodal distribution. We acknowledge this limitation as this scenario was not examined in the current simulation study.

Two useful practical recommendations can be extracted from the current study. As previously mentioned, no standard method exist for determining appropriate tuning parameter values that ensure the 20-50% acceptance rate needed for the MCMC procedure. Although trial and error also is used in this study, our findings provide a general guideline for the selection of tuning parameter values. We find that tuning parameter values 1.00, 0.70, 0.50, and 0.30 appear to work well for sample sizes 50, 100, 200, and 500, respectively. Additionally, our study results are consistent with findings from Polit and Beck (2006) regarding the number of subject experts needed to establish content validity. Across three types of expert biases, results show that having more than six experts does not contribute to any additional gain in the efficiency of OBID. With highly biased experts, three experts appear to be sufficient for establishing content validity.

An implication from this study is that a hierarchical model can be considered in the future to incorporate the individual effect of content experts, as the scores experts assigned from item to item are likely to be correlated. In addition, the development of the user-friendly BID software can be used to guide the development of the OBID software, where multi-factor models can be evaluated, as it is common in many "long-form" questionnaires to encompass several constructs of interest. It is our ultimate goal to extend the application capability of OBID and present it as an efficient and reliable method for researchers and clinicians in future PROMs development.

2.5 Conclusions

In this study, the efficiency of OBID is evaluated by comparing its performance to classical instrument development performance using actual and simulation data. This study successfully demonstrated that the OBID approach is more efficient than the classical approach

when the sample size is small. OBID promises an efficient and reliable method for researchers and clinicians in future PROMs development for small populations or rare diseases.

Chapter Three

A Novel Method for Expediting the Development of Patient-Reported Outcome Measures and an Evaluation Across Several Populations

Lili Garrard, Larry R. Price, Marjorie J. Bott, and Byron J. Gajewski

(Submitted to *Applied Psychological Measurement*)

Abstract

Item response theory (IRT) models provide an appropriate alternative to the classical ordinal confirmatory factor analysis (CFA) during the development of patient-reported outcome measures (PROMs). Current literature has identified the assessment of IRT model fit as both challenging and underdeveloped (Sinharay & Johnson, 2003; Sinharay et al., 2006). This study evaluates the performance of Ordinal Bayesian Instrument Development (OBID), a Bayesian IRT model with a probit link function approach, through applications in two breast cancer-related instrument development studies. The primary focus is to investigate an appropriate method for comparing Bayesian IRT models in PROMs development. An exact Bayesian leave-one-out cross-validation (LOO-CV) approach (Vehtari & Lampinen, 2002) is implemented to assess prior selection for the item discrimination parameter in the IRT model and subject content experts' bias toward the estimation of item-to-domain correlations. Results support the utilization of content subject experts' information in assessing construct validity for small sample sizes. However, the incorporation of subject experts' content information in the OBID approach can be sensitive to the level of expertise of the recruited experts. More stringent efforts need to be invested in the appropriate selection of subject experts to use efficiently the OBID approach and reduce potential bias during PROMs development.

Keywords: OBID, Bayesian leave-one-out cross-validation, Bayesian IRT, Bayesian model comparison, patient-reported outcome measures, PROMs

3.1 Introduction

The famous statistician George E. P. Box once said: “all models are wrong, but some are useful.” No statistical model is adequate in capturing all mechanisms presented in real data. Researchers often build a few candidate models and seek to select the most useful one for a given problem. The process of model comparison and selection requires rigorous model checking or assessment that is an integral part of any statistical analysis. In the development of psychometric instruments, apart from reliability, establishing evidence of validity is essential to ensuring an instrument’s psychometric integrity. Developing an evidence-based argument that scores are accurate for their intended use requires acquiring data specific to content, construct, and predictive aspects (Nunnally & Bernstein, 1994). Historically, validity has been presented as three distinct but related components—content, criterion and construct. Today validity is viewed as a unitary concept (AERA, APA, & NCME, 2014) where propositions for test score interpretation and use are supported by evidence unique to the measurement goal. Although developing a comprehensive picture of score validity often includes content and predictive components, construct validity receives the most attention from a statistical modeling perspective. The reason that construct validity receives the most attention is because any score validity argument is impossible to make without evidence that the construct is relevant to the proposed interpretation and use of the scores.

Two approaches can be implemented to establish evidence of construct validity. When the participant sample size is adequately large, classical (i.e., frequentist) confirmatory factor analysis (CFA) is fairly reliable and easy to implement via statistical software such as Mplus (L. K. Muthén & Muthén, 1998-2012) or the free R package *lavaan* (Rosseel, 2012). Bayesian approach becomes advantageous when classical CFA is challenged by small sample size

(Gajewski et al., 2013; Garrard, Price, Bott, & Gajewski, 2015; Jiang et al., 2014), which may result in model convergence issues and unreliable parameter estimates.

An emerging topic in recent literature focuses on the development of patient-reported outcome measures (PROMs) or patient-reported outcome (PRO) instruments that often are designed as survey instruments with ordinal response options. PROMs have gained increasing public awareness in promoting patient-centered care, one of the most important driving forces behind the current U.S. health care. For instance, the pharmaceutical industry is required by the U. S. Department of Health and Human Services (DHHS) Food and Drug Administration (FDA) to submit evidence collected through PRO instruments in support of labeling claims. Detailed industry guidelines are provided by the FDA to assist pharmaceutical companies regarding the psychometric evaluation of any new or adapted PRO instruments (FDA, 2009).

Ordinal or binary (a special type of ordinal data) patient responses often are collected from PROMs that require a different modeling approach when compared to the classical CFA (e.g., normality assumption) for assessing the instrument's construct validity. Literature has shown that the classical CFA model is equivalent to a two-parameter item response theory (IRT) model with a probit link function, when all the items on an instrument are ordinal (Garrard et al., 2015; Johnson & Albert, 1999; Quinn, 2004). However, assessing the fit of IRT models remains a challenging and underdeveloped area in the literature (Sinharay & Johnson, 2003; Sinharay et al., 2006). This paper extends the current literature by focusing the discussions around IRT model comparison using the Bayesian framework.

In general, there are several ways that the fit of Bayesian models can be evaluated. One popular method is posterior predictive model checking (PPMC; Rubin, 1984), which is closely related to classical goodness-of-fit tests (Gelman, Meng, & Stern, 1996; Sinharay & Johnson,

2003). Other methods include graphical posterior predictive checks, assessing the posterior predictive *p*-value, and/or the utilization of Bayes factors (Gelman, Hwang, & Vehtari, 2014). However, as pointed out by Gelman et al. (2014), when the objective is to compare models, the predictive model accuracy needs to be estimated. Cross-validation and information criteria measures are commonly used for Bayesian model comparison (Gelman et al., 2014; Vehtari & Lampinen, 2002; Vehtari & Ojanen, 2012). Information criteria typically are defined as deviance measures and represented by some variations of the log-likelihood or log predictive density. Stone (1977) has shown the asymptotic equivalency between the two approaches such that information criteria can be viewed as approximations to various types of cross-validation (Gelman et al., 2014).

Despite several common criticisms, deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2014; Spiegelhalter, Best, Carlin, & van der Linde, 2002) remains a popular choice in the Bayesian literature and easily can be computed via the software WinBUGS (Lunn et al., 2000). Viewed analogously to the well-known Akaike information criterion (AIC; Akaike, 1973), DIC is considered as another pointwise measure for conditioning on the poster mean, whereas AIC conditions on the maximum likelihood estimator (MLE). A more fully Bayesian approach, known as WAIC (widely applicable or Watanabe-Akaike information criterion), recently has been proposed by Watanabe (2010). WAIC is considered more appealing than AIC and DIC as it not only conditions on the entire posterior distribution, but also works well with hierarchical and mixture structure models (Gelman et al., 2014). Among other cross-validation methods for evaluating out-of-sample prediction performance, Bayesian leave-one-out cross-validation (LOO-CV) has been shown to be asymptotically equivalent to WAIC (Watanabe, 2010) and more applicable to problems with small *n*.

Although both WAIC and Bayesian LOO-CV exhibit appealing properties, they are applied less in practice as Bayesian cross-validation approaches can become very computational intensive due to Markov chain Monte Carlo (MCMC) simulation for all validation units. Several approximation approaches have been proposed in the literature for Bayesian LOO-CV, such as importance sampling (IS; Gelfand, Dey, & Chang, 1992), expectation propagation and Laplace approximation (Vehtari, Tolvanen, Mononen, & Winther, 2014), Bayesian *K*-fold cross-validation (Vehtari, Gelman, & Gabry, 2015), and a more recent Pareto smoothed importance sampling (PSIS) approach for regularizing importance weights (Vehtari & Gelman, 2015). The new PSIS approach has been incorporated into the R package *loo* (Vehtari et al., 2015).

Within the context of latent variable modeling, excellent research recently has been conducted that approximates Bayesian LOO-CV for Gaussian latent variable models (Li, Qiu, Zhang, & Feng, 2014; Vehtari et al., 2014). Interested readers may refer to the above references for details on the various approximation methods. As previously mentioned, common data collected from PROMs are ordinal in nature, which calls for an extension of the Gaussian model method to ordinal models (i.e., IRT models). Yet, there is a lack of Bayesian LOO-CV approximation with ordinal models in the current literature (A. Vehtari, personal communication, July 20, 2015). In addition, Bayesian model comparison should be evaluated from the perspective of prior selection for the IRT model parameters. The choice of prior distribution is relevant to posterior parameter inferences and model predictions when data are sparse (Gelman et al., 2014).

During the development of PROMS, when the sample size for an intended patient population is small (e.g., cases of rare disease), a novel method called Ordinal Bayesian Instrument Development (OBID) recently has been proposed to overcome the small sample size

challenge and appropriately model the participants' ordinal responses (Garrard et al., 2015). OBID is developed within a Bayesian two-parameter IRT with a probit link modeling framework. *Prior* distributions derived from content subject experts' data (for establishing content validity of the instrument) are updated with participants' data to obtain a *posterior* distribution for the IRT model parameters.

The work in this paper is motivated by prior research on Bayesian LOO-CV for Gaussian latent variable models and the need for having an appropriate method for comparing Bayesian IRT models in PROMs development. The OBID approach is evaluated through real data applications and the specific aims include: a) comparing the OBID models with both informative and flat priors using exact Bayesian LOO-CV, and b) assessing subject content experts' bias through an exact CV information criterion (IC) measure. All real data used in the current study were collected for prior research purposes and provided to the authors in a de-identified fashion. Thus, this study was determined as non-human subject research by a Midwestern academic medical center Internal Review Board (IRB).

3.2 Methodology

Since the main objective of this paper is to evaluate further the OBID approach through Bayesian model comparison using real data applications, we first will provide a brief review of the OBID participant model and how an exact Bayesian LOO-CV can be applied to the scenarios that will be discussed in the current study.

3.2.1 OBID Participant Model

OBID is an ordinal CFA-based approach under the Bayesian probabilistic framework (Garrard et al., 2015). As mentioned in the introduction, the IRT model can be viewed as the

ordinal version of classical CFA. Continuing the notations from Garrard *et al.*, a two-parameter IRT model with the probit link is expressed by

$$y_{ij} = c \text{ if } y_{ij}^{*} \in \left(T_{j(c-1),} T_{jc}\right]; \;\; i = 1, \ldots, N, \; j = 1, \; \ldots, \; P, \; c = 1, \ldots, \; C_j \tag{3.1}$$

$$y_{ij}^{*} = \alpha_j + \lambda_j f_i + \varepsilon_{ij}; \;\; f_i \sim \; N(0,1), \varepsilon_{ij} \sim N(0,1), i = 1, \ldots, \; N, \; j = 1, \; \ldots, \; P, \tag{3.2}$$

where y_{ij} represents the *i*th participant's response to the *j*th item; and C_j is the number of response options for the *j*th item. The ordinal response y_{ij} is related to a continuous latent variable y_{ij}^{*}, through a set of ordered cut-points T_{jc}, on y_{ij}^{*}. The two item-specific parameters are α_j, the negative difficulty parameter for the *j*th item, and λ_j, the discrimination parameter for item *j*. The latent ability variable f_i is constrained to follow a standard normal distribution with ε_{ij} being the measurement error. The model further can be interpreted such that the probability of a particular response option being endorsed depends on the probability that y_{ij}^{*} falls within an interval defined by the cut-points.

Under the local independence or conditional item independence assumption (Price, in press), the likelihood for the underlying continuous latent variable y_{ij}^{*} is

$$L(\boldsymbol{y}^{*} | \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{f}) = \prod_{i=1}^{N} \prod_{j=1}^{P} N(y_{ij}^{*} | \alpha_j + \lambda_j f_i \; , \; 1). \tag{3.3}$$

In the unidimensional (i.e., single-factor) OBID approach, the prior distribution of the item discrimination parameter λ_j is specified using content validity information from subject experts (i.e., item relevancy ratings/latent item-to-domain correlations; informative prior). Suppose x_{jk} represents the *k*th expert's relevancy rating for the *j*th item; and ρ_{jk} represents the same *k*th expert's latent item-to-domain correlation for the *j*th item. The common four-point relevancy scale used by the experts (i.e., 1 = "not relevant", 2 = "somewhat relevant", 3 = "quite relevant", 4 = "highly relevant") and the latent correlation scale can be related to each other through either

an equally-spaced (i.e., $0.00 \leq \rho_{jk} < 0.25$; $0.25 \leq \rho_{jk} < 0.50$; $0.50 \leq \rho_{jk} < 0.75$; and $0.75 \leq \rho_{jk} < 1.00$, respectively) or unequally-spaced (i.e., $0.00 \leq \rho_{jk} < 0.10$; $0.10 \leq \rho_{jk} < 0.30$; $0.30 \leq \rho_{jk} < 0.50$; and $0.50 \leq \rho_{jk} < 1.00$, respectively) transformation. Findings by Gajewski et al. (2012) suggest that for a panel of individuals with moderate level of expertise in the area of interest, the equally-spaced transformation might be more appropriate. Interested readers are referred to Gajewski et al. (2012), Gajewski et al. (2013), Jiang et al. (2014), and Garrard et al. (2015) for additional background and details on the OBID approach.

3.2.2 Bayesian Leave-one-out Cross-validation (LOO-CV)

Bayesian cross-validation is a common method used to evaluate out-of-sample prediction performance and compare models. The idea behind cross-validation is quite intuitive and our description of the method intentionally is kept consistent with the work by Gelman et al. (2014) and Li et al. (2014). First, the full dataset repeatedly can be partitioned into a holdout data $\boldsymbol{y}_i$ and a training data $\boldsymbol{y}_{-i}$. Since the focus is on LOO-CV, the holdout dataset in our application will simply be a single participant's responses to all items on an instrument. Second, the model is fitted to the training data $\boldsymbol{y}_{-i}$, yielding the posterior distribution $P_{post(-i)}(\boldsymbol{\theta}, \boldsymbol{f}|\boldsymbol{y}_{-i})$ of the model parameters $\boldsymbol{\theta}$ and the latent variable $\boldsymbol{f}$, all denoted in the general notation format. Third, the posterior predictive density of the holdout data $\boldsymbol{y}_i$, conditioning on the training data, can be computed by specifying an evaluation function $a(\boldsymbol{y}_i, \boldsymbol{\theta}, \boldsymbol{f}_i)$ that measures certain goodness-of-fit of the prediction to the actual holdout observation $\boldsymbol{y}_i$.

Following the work by Li et al. (2014), the CV posterior predictive evaluation is defined as the expectation of the evaluation function $a(\boldsymbol{y}_i, \boldsymbol{\theta}, \boldsymbol{f}_i)$ with respect to the posterior distribution of the parameters, conditioning on the training data that can be expressed by

$$E_{post(-i)}\{a(\boldsymbol{y}_i, \boldsymbol{\theta}, \boldsymbol{f}_i)\} = \int a(\boldsymbol{y}_i, \boldsymbol{\theta}, \boldsymbol{f}_i)\, P_{post(-i)}(\boldsymbol{\theta}, \boldsymbol{f}|\boldsymbol{y}_{-i}) d\boldsymbol{\theta} d\boldsymbol{f}. \quad (3.4)$$

Suppose we let the evaluation function be the value of the predictive density function at the actual holdout observation $\boldsymbol{y}_i$ (i.e., $a(\boldsymbol{y}_i, \boldsymbol{\theta}, \boldsymbol{f}_i) = P_{pred}(\boldsymbol{y}_i|\boldsymbol{\theta}, \boldsymbol{f}_i)$), the CV posterior predictive evaluation (Equation 3.4) becomes the CV posterior predictive density that can be approximated by averaging the predictive densities at the actual holdout observation $\boldsymbol{y}_i$, across all MCMC draws from $P_{post(-i)}(\boldsymbol{\theta}, \boldsymbol{f}|\boldsymbol{y}_{-i})$. The CV posterior predictive density is expressed by

$$P_{pred}(\boldsymbol{y}_i|\boldsymbol{y}_{-i}) = \int P_{pred}(\boldsymbol{y}_i|\boldsymbol{\theta}, \boldsymbol{f}_i)\, P_{post(-i)}(\boldsymbol{\theta}, \boldsymbol{f}|\boldsymbol{y}_{-i}) d\boldsymbol{\theta} d\boldsymbol{f} \quad (3.5)$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} P_{pred}^{s}(\boldsymbol{y}_i|\boldsymbol{\theta}^s, \boldsymbol{f}_i^s). \quad (3.6)$$

As previously introduced, the participant model (Equations 3.1 and 3.2) in the current study is a single-factor two-parameter IRT model, where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\lambda})$. For each *s*th MCMC posterior draw, estimates of the negative item difficulty parameter α_j^s, the item discrimination parameter λ_j^s, and the cut-points T_{jc}^s on y_{ij}^*, can be obtained for each item on the instrument. In the current model, the latent variable f_i^s requires no updating, as the MCMC draws come from the prior distribution. Then the predictive density at the actual holdout observation $\boldsymbol{y}_i$ at each MCMC iteration (i.e., $P_{pred}^{s}(\boldsymbol{y}_i|\boldsymbol{\alpha}^s, \boldsymbol{\lambda}^s)$) can be computed as the multivariate normal distribution function evaluated on the intervals defined by the cut-points of each item that is expressed by

$$P_{pred}^{s}(\boldsymbol{y}_i|\boldsymbol{\alpha}^s, \boldsymbol{\lambda}^s) = \int_{\boldsymbol{T_{c-1}^s}}^{\boldsymbol{T_c^s}} MVN\left(\boldsymbol{y}_i^*|\boldsymbol{\alpha}^s + \boldsymbol{\lambda}^s f_i^s, \boldsymbol{I}\right) d\boldsymbol{y}_i^*. \quad (3.7)$$

Finally, the CV information criterion (CVIC; Li et al., 2014) is computed by -2 times the sum of the log of the CV posterior predictive density, over all validation units. The model with the smaller CVIC value is preferred.

To demonstrate the computation of the CV posterior predictive density $P_{pred}(\boldsymbol{y}_i|\boldsymbol{y}_{-i})$ (Equations 3.5 and 3.6), we will use a hypothetical three-item instrument with binary response

options (i.e., 1/0 or correct/incorrect). Suppose the holdout data $\boldsymbol{y}_i$ represent the *i*th subject's responses to the three items, where $\boldsymbol{y}_i = (0,1,1)$. For items with binary response options, the single cut-point on the underlying continuous latent variable $y_{ij}^*$is zero. The binary response y_{ij} is related to the latent variable y_{ij}^* through the following function:

$$y_{ij} = \begin{cases} 0 & if\ y_{ij}^* \in (-\infty, 0] \\ 1 & if\ y_{ij}^* \in (0, \infty] \end{cases}. \tag{3.8}$$

The holdout data $\boldsymbol{y}_i$ can be used to determine the corresponding set of cut-points needed for each of the three items. For instance, the *i*th participant's actual response for the first item is 0; therefore the set of cut-points used will be $(-\infty, 0]$. At each *s*th MCMC iteration, we can specify the set of cut-points $(\boldsymbol{T}_{\boldsymbol{c-1}}^{\boldsymbol{s}}, \boldsymbol{T}_{\boldsymbol{c}}^{\boldsymbol{s}}]$, where $\boldsymbol{T}_{\boldsymbol{c-1}}^{\boldsymbol{s}} = (-\infty, 0, 0)$ and $\boldsymbol{T}_{\boldsymbol{c}}^{\boldsymbol{s}} = (0, \infty, \infty)$**.** The predictive density $P_{pred}^{s}(\boldsymbol{y}_i|\boldsymbol{\alpha}^s, \boldsymbol{\lambda}^s)$ can be computed by evaluating the three-dimensional multivariate normal distribution function on the intervals defined by these cut-points, using the R function *pmvnorm* (Genz et al., 2015; R Core Team, 2015). Finally, the CV posterior predictive density $P_{pred}(\boldsymbol{y}_i|\boldsymbol{y}_{-i})$ at the actual holdout observation $\boldsymbol{y}_i$ is approximated by averaging across all MCMC iterations. The rest of the computations in this paper are performed using the R package *MCMCpack* (Martin et al., 2011; R Core Team, 2015) and the software WinBUGS (Lunn et al., 2000).

3.3 Real Data Applications

In this section, data collected from two breast cancer-related instrument development studies will be described and analyzed using the recently proposed OBID approach. An exact Bayesian LOO-CV is applied to compare the choice of prior for the item discrimination

parameter λ_j (see Equation 3.2), and to assess subject experts' bias toward the item-to-domain correlation (or item relevancy), under both equally-spaced and unequally-spaced transformations.

3.3.1 PAMS-Short Form Satisfaction Survey

PAMS background. Breast cancer related death ranks second among cancer deaths for women in the U. S. (DHHS, 2011). Routine utilization of mammography is the most widely recommended method for breast cancer screening and offers patients a chance of early detection that is critical for overall survival. However, potential factors such as prior experiences and satisfaction with mammography influence patients' decision on using mammography on a regular basis. The patient assessment of mammography services (PAMS) satisfaction survey was developed due to the lack of mammography-specific satisfaction assessments (Engelman et al., 2010; Engelman et al., in review). The full PAMS survey consists of four-factors with 20 items and the PAMS-Short Form is a single factor with seven items. Items on the full survey are designed with scales ranging from two to six response categories. The seven short-form items can be rated on a five-point, Likert-type scale (i.e., 1 = "poor", 2 = "fair", 3 = "good", 4 = "very good", and "5 = excellent.")

PAMS experts and participants. Six subject experts were consulted and instructed to rate the relevancy of each item (ranging from 1 = "not relevant" to 4 = "highly relevant") to the domain of interest. The recruited experts consist of individuals who have published or worked in some type of breast cancer research, including several physicians (Ndikum-Moffor et al., in review). In addition, participant data were collected from female patients to establish construct validity of the PAMS-Short Form instrument. Complete data (i.e., participants responded to all items) are used for the current study. The patients represented four different ethnicity

backgrounds: Hispanic (n = 36), Non-Hispanic white (n = 2,768), Black (n = 34), and American Indian (n = 287).

PAMS CV – prior selection. For the current study, analyses focused on the Hispanic, Black, and American Indian populations. First, distribution of response options (potential range = 1 to 5) from the raw participant data were examined. Very few respondents selected poor to good response options; thus a decision was made to collapse some of the response categories. Potential loss of information due to scale reduction is acknowledged; however this decision should not greatly affect the general trend in the data. For Hispanic and Black data, the five-point scale is reduced to a three-point scale by collapsing poor, fair, and good response options into one category; and poor to fair response options are collapsed into one category for the American Indian data, turning the scale into a four-point scale.

The OBID approach promotes the incorporation of content experts' information (when appropriate) for the item discrimination parameter λ_j. In the absence of subject experts or an appropriate prior reference data (Garrard et al., 2015), a flat prior (i.e., $\lambda_j \sim N(0, 4)$) can be used to fit the model and obtain parameter estimates. Furthermore, an exact Bayesian LOO-CV is applied to compare the choice of using a flat prior versus an informative prior (under both transformations). The CVIC values for the flat prior, the equally-spaced transformation prior, and the unequally-spaced transformation prior are 589.93, 503.68, and 482.87 for Hispanic; 598.06, 525.01, and 485.83 for Black; and 4112.87, 4042.25, and 4054.88 for American Indian, respectively. Across all three patient populations, both types of informative prior produce smaller CVIC values than that of the flat prior. Since models with smaller CVIC values are preferred, results indicate that the unequally-spaced transformation models are preferred for the Hispanic

and Black populations; whereas the equally-spaced model appears to be slightly better than the unequally-spaced model for the American Indian population.

PAMS CV – expert bias. It is beneficial to assess experts' bias toward the item-to-domain correlation (or item relevancy), especially for smaller sample sizes. Figure 3.1 displays the CVIC value for each selected number of experts *K*. Recall that the total number of experts for the PAMS study is six. The CVIC is calculated by both randomly selecting one to five experts

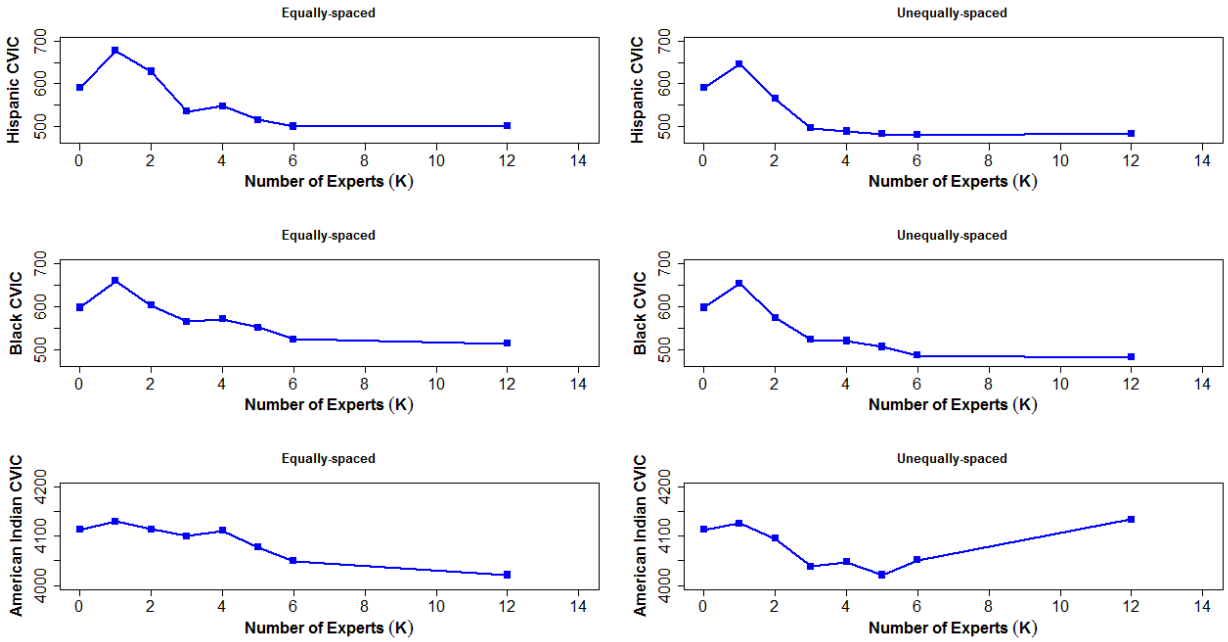


Figure 3.1. PAMS expert bias comparison under both equally-spaced (left panel) and unequally spaced (right panel) transformations.
Note. $K = 0$ implies flat priors.

from the pool of six experts and artificially inflating the prior sample size to represent information from 12 experts. *K*=0 implies the use of flat prior that is added to the plots for comparison purposes. As the number of experts increases, the majority of CVIC values under the unequally-spaced transformation are smaller than that of the equally-spaced transformation. The selected experts appear to be less biased for both Hispanic and Black populations. However, the

same group of experts is slightly more biased for the American Indian population. The CVIC value sharply increases after five experts for the unequally-spaced transformation; whereas the CVIC value continues to decrease for the equally-spaced transformation. In addition, all three equally-spaced transformation plots indicate that six experts are adequate, which is consistent with the suggestion in the current literature (Polit & Beck, 2006).

3.3.2 NLit-BCa Study

NLit-BCa background. The Nutrition Literacy Assessment Instrument (NLit) was originally developed by H. Gibbs and Chapman-Novakofski (2013) to assess nutrition literacy. Pilot work conducted by H. D. Gibbs et al. (2015) initiated the creation of the Nutrition Literacy Assessment Instrument for Breast Cancer (NLit-BCa), which is adapted from the original NLit for female breast cancer survivors, as there is a lack of nutrition literacy instrument for this specific patient population. The adapted NLit-BCa consists of six individual domains with 75 items. A larger validity study is currently in process to evaluate further the NLit-BCa instrument (H. Gibbs, personal communication, August 25, 2015). Considering the item revisions and/or deletions based on content experts' review, four domains with 39 items (i.e., ten macronutrients (Macro) items, nine household food measurement (HFM) items, ten food label and numeracy (FLN) items, and ten consumer skills (CS) items) are deemed appropriate for analysis in the current study. Items are designed with either three or four response options; and all participant responses are further classified as 0 "incorrect" and 1 "correct" based on an answer key provided by the instrument developers.

***NLit-BCa experts and participants*.** Four nutrition experts were consulted for the larger validation study and rated the relevancy of each item on the 75-item instrument. The recruited

experts consist of individuals who have published expertise in cancer nutrition. Since the larger validation study is on-going, the participant data will come from the pilot work. Data originally were collected from two groups of participants: the weight loss intervention group and the non-intervention group. Due to data sparsity concerns, complete data from 71 patients are used after combining the two groups (n = 25 and 46 for the intervention and the non-intervention groups, respectively).

NLit-BCa CV – prior selection. Prior to analysis, a decision was made to exclude both item 3 from the macronutrients domain (Macro03) and item 2 from the food label and numeracy domain (FLN02) to avoid potential issues for the LOO-CV analyses. Only one respondent answered Macro03 incorrectly, and everyone correctly answered FLN02. Thus the total number of items was 37. The choice of flat prior versus an informative prior under both transformations was compared using exact Bayesian LOO-CV. The CVIC values for the flat prior, equally-spaced transformation prior, and the unequally-spaced transformation prior are 504.10, 506.64, and 507.36 for Macro; 927.94, 947.59, and 941.58 for HFM; 633.84, 660.89, and 664.57 for FLN; and 716.07, 720.99, and 719.55 for CS, respectively. Across all four domains, the flat prior produces smaller CVIC values than both types of informative prior; however, the differences in CVIC values are much smaller for the consumer skills (CS) domain.

NLit-BCa CV – expert bias. Results from the prior selection analysis seem to suggest that the content experts are more biased toward the item-to-domain correlations for all four domains. Figure 3.2 shows the CVIC value for each selected number of experts K. Similar to the PAMS study, the CVIC is calculated by both randomly selecting one to three experts from the pool of four experts and artificially inflating the prior sample size to represent information from

eight experts. The use of flat prior again is indicated by *K*=0. For the Macro domain, the CVIC value continues to decrease under the equally-spaced transformation prior after four experts,

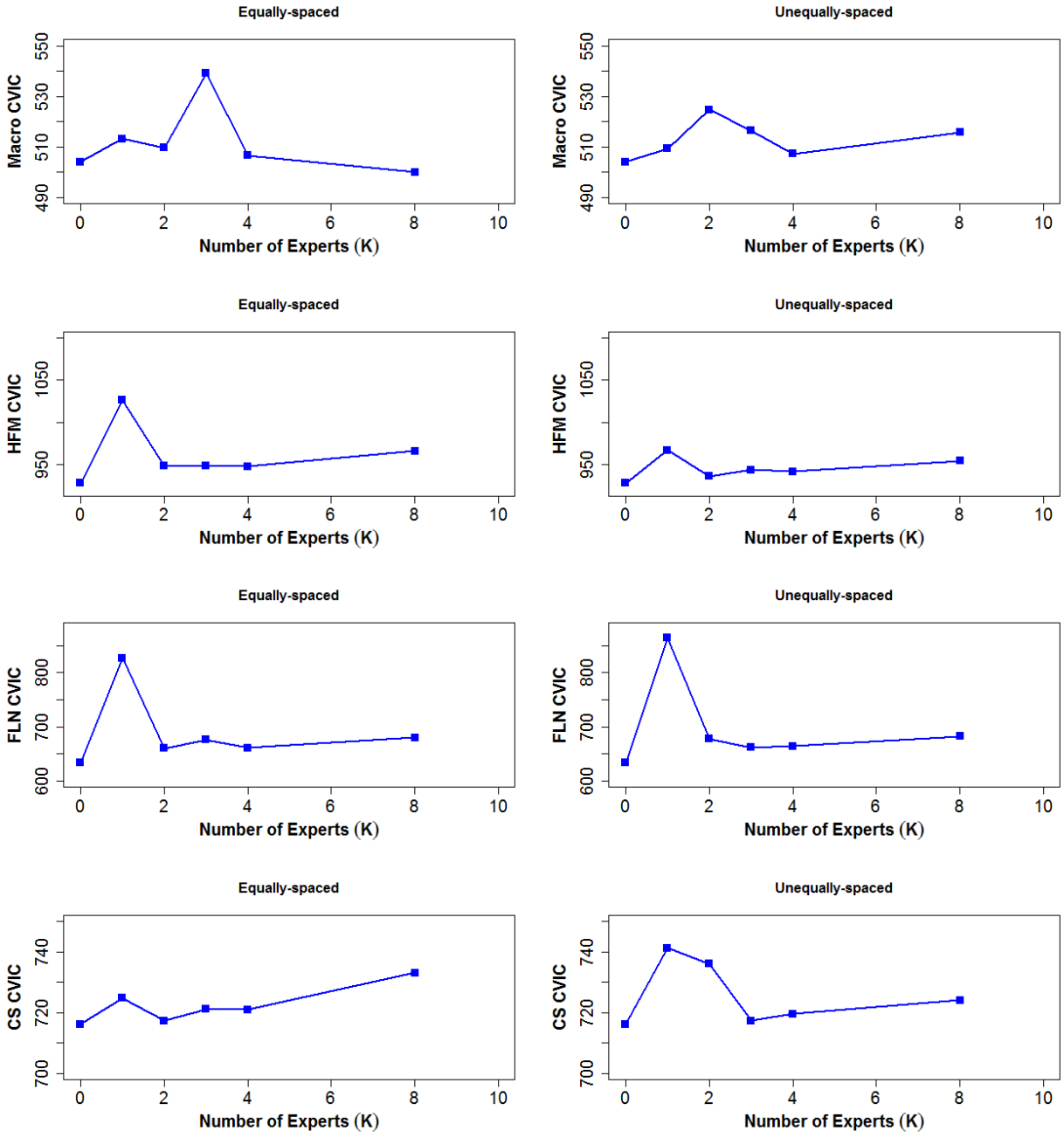


Figure 3.2. NLit-BCa expert bias comparison under both equally-spaced (left panel) and unequally spaced (right panel) transformations.
Note. K = 0 implies flat priors; Macro = macronutrients; HFM = household food measurement; FLN = food label and numeracy; CS = consumer skills.

where the opposite is observed with the unequally-spaced transformation prior. No huge differences in the CVIC values are observed among two to four experts for both household food (HFM) and food label and numeracy (FLN) domains, under both transformations. For the CS domain, apart from the flat prior model, two experts produce the smallest CVIC value under the equally-spaced transformation, whereas the smallest CVIC value occurs with three experts. Overall, the recruited experts seem to be more biased toward the relevancy ratings on the items, across all four domains.

3.4 Discussion

The current study evaluates the performance of OBID through applications in two breast cancer-related instrument development studies. The primary focus is to investigate an exact Bayesian LOO-CV approach for comparing Bayesian IRT models in PROMs development. Six subject experts are consulted in the PAMS-short form study for four different patient populations. Among the three populations investigated in the current study, the use of an informative prior (i.e., incorporating experts' information) has shown to be superior to the use of a flat prior. One interesting observation arises from the original focus of the six content experts, as the experts were originally recruited with the purpose of validating the PAMS instrument for the American Indian women. Results from the PAMS study indicate that the experts are less biased for both Hispanic and Black populations, which supports the appropriate utilization of experts' information to form a "general prior" as suggested by Garrard et al. (2015). The experts appear to be slightly more biased for the American Indian population despite their original focus. Although findings suggest that five experts would be sufficient, the use of six experts does not pose any substantial concerns for the purpose of instrument validation. For demonstration

purposes, each participant's expected responses for all items are calculated using parameter estimates from the posterior draws. Figure S3.1 in the appendix displays a comparison between the original data and the expected data for the proportion of Hispanic participants selecting each of the three response options, across all seven items. The Black and American Indian population comparisons are not reported here. Overall results indicate that incorporating information from the six selected subject experts is appropriate for the construct validity analysis in the PAMS study.

Findings from the NLit-BCa study present more complexity as the current study suggest the use of a flat prior as opposed to an informative prior. Among the four domains examined, only the FLN domain CVIC results slightly support incorporating experts' information. The four selected experts appear to hold more biased opinions regarding the item-to-domain correlations for the items in all domains. Although four experts were recruited, results have shown that even two to three experts would be sufficient. One thing worth noting is that the design of the NLit-BCa study differs from the PAMS study. The PAMS items are more subjective (i.e., eliciting satisfaction); whereas, the NLit-BCa items have a distinct correct answer. Nonetheless, despite the seemingly "opposite" results from the NLit-BCa study, the importance of appropriate prior selection and expert bias evaluation has been demonstrated for the OBID approach.

One limitation of the current study is associated with the selection of subject content experts, which remains an important yet challenging aspect in the development of psychometric instruments (Grant & Davis, 1997; Lynn, 1986). Apart from unidimensional instruments, the subject experts often are asked to rate items from multiple domains. It usually is assumed that the experts have expertise in all areas of interest. The current study assumes that the content validity has been assessed thoroughly for both instruments. Thus the focus is entirely on model selection

during the construct validity phase of the instrument development. Yet, based on findings from the current study, subject experts' bias may hinder the efficient utilization of experts' information in the recently proposed OBID approach. Another limitation comes from the primary focus on using an exact Bayesian LOO-CV approach to compare different IRT models. As mentioned in the introduction, several methods can be used to help assess and compare Bayesian models. The OBID approach certainly can be evaluated further via other established approaches, such as *K*-fold CV. The literature suggests that *K*-fold CV has advantage over LOO-CV due to a bias-variance trade-off. The test error rate from *K*-fold CV tends to have smaller variance than that of the LOO-CV approach (Breiman & Spector, 1992; James, Witten, & Hastie, 2014; Kohavi, 1995). The third limitation can be viewed as a constraint associated with using the R package *MCMCpack,* as normal priors are required for the IRT model parameters. Future work can consider other types of prior distributions.

An implication from the current study is the selection of an appropriate tuning parameter to ensure 20-50% acceptance rate during the MCMC procedure. Simulation results from Garrard et al. (2015) have shown an inverse relationship between the tuning parameter and the sample size. Although not discussed in the main text of the paper, based on sample size information from 11 real data sets and the four simulation data sets from Garrard *et al.*, a power function is fitted for the tuning parameter *t* as a function of sample size *n*, i.e. $t = 11.947n^{-.544}$ with $R^2 = .84$. This formula should be further refined as more data sets become available.

Additional future work may involve a more thorough evaluation of the equally-spaced and unequally-spaced transformations in other real applications and an approximation to the Bayesian LOO-CV for ordinal latent variable models. In addition, more skewed participant data structure and other prior distributions for the OBID subject experts' model need to be evaluated

through simulation. The simulation study by Garrard et al. (2015) considers a more balanced participant data structure and that the experts' item ratings follow a normal distribution. For instruments with more subjective response scales (e.g., satisfaction), the participants tend to select more positive response options. The experts can also potentially disagree with each other regarding the relevancy of proposed items.

Chapter Four

Reliability and Validity of the NDNQI® Injury Falls Measure

Lili Garrard, Diane K. Boyle, Michael Simon, Nancy Dunton, and Byron J. Gajewski

Garrard, L., Boyle, D. K., Simon, M., Dunton, N., & Gajewski, B. (2014). Reliability and validity of the NDNQI® Injury Falls Measure. *Western Journal of Nursing Research*, 0193945914542851.

Abstract

Although remarkable efforts have been made to improve patient fall reporting through the utilization of standardized definitions, injury falls reporting rarely has been examined. This study used an overall intra-class correlation coefficient (ICC) estimate and factor analysis to assess the reliability and validity of the National Database of Nursing Quality Indicators® (NDNQI®) falls with injury measure. Data were collected from an online Fall Injury Level Survey that was administered to 1,159 NDNQI site coordinators (39.7% response rate; 91% registered nurses [RNs]). Estimated overall ICC was .85. Exploratory factor analysis (EFA) with a Promax rotation (root mean square error of approximation [RMSEA] = 0.053) identified three latent factors: No Injury, Minor Injury, and Moderate/Major Injuries. Final confirmatory factor analysis (CFA) assessment (comparative fit index [CFI] = 0.914, Tucker Lewis Index [TLI] = 0.910, RMSEA = 0.048) confirmed an acceptable model fit. Results provided strong evidence that the NDNQI falls with injury measure is reliable and valid in supporting hospitals' fall prevention efforts and future injurious falls research.

Keywords: injury falls, fall injury levels, reliability, NDNQI

4.1 Introduction

Falls are common adverse events experienced by patients in hospitals and continue to pose challenges to health care quality. Fall reduction is identified as a patient safety priority in the United States (National Priorities Partnership, 2011). Approximately 30% of falls result in injury, particularly among older adults (Shorr et al., 2008). Injuries from falls burden hospitals and patients with increased costs due to longer lengths of stay and additional patient care costs (Currie, 2008). For older adults, the direct and indirect cost of injuries associated with falls is projected to reach U.S. $54.9 billion (in year 2007 dollars) annually by 2020 (Centers for Disease Control and Prevention, 2013; Englander, Hodson, & Terregrossa, 1996). In an effort to promote patient safety, the National Quality Forum (NQF; 2011) named "patient death or serious injury associated with a fall while being cared for in a healthcare setting" (p. 9) as one of the health care Serious Reportable Events (SREs). Similarly, the Centers for Medicare & Medicaid Services (CMS) identified hospital falls and resulting trauma as one of the preventable Hospital-Acquired Conditions (HAC). Additional costs associated with HAC are no longer covered by Medicare for hospitals participating in the Inpatient Prospective Payment System (IPPS; CMS, 2012; Inouye, Brown, & Tinetti, 2009).

4.1.1 National Database of Nursing Quality Indicators® (NDNQI®) Fall and Falls With Injury Measures

NQF established a national framework to evaluate health care quality measurement and reporting. NQF's goals are to increase public awareness in quality performance, establish incentives for performance improvement, and provide national benchmarks (NQF, 2002; Simon, Klaus, Gajewski, & Dunton, 2013) . Both patient fall and falls with injury have been endorsed by

the NQF as national consensus measures since 2004 (NQF, 2004). The American Nurses Association (ANA), serving as the NQF steward for both measures, commissioned the NDNQI to conduct separate studies to assess the reliability of each fall measure in part to support their successful NQF re-endorsement in 2013 (NQF, 2013a, 2013b). NDNQI was established in 1998 by ANA to monitor nurse-sensitive quality indicators that are essential for patient safety and quality improvements in hospitals (Montalvo, 2007). NDNQI is a quality database that collects and evaluates unit-specific nurse-sensitive data from over 2,000 U.S. and international hospitals. Member hospitals of NDNQI benefit from regular reporting of nursing quality measures and various national comparison data that were shown to be helpful in quality improvement.

The NDNQI patient fall reliability study was conducted by Simon and colleagues (2013) to examine the agreement of fall classifications among staff in U.S. hospitals (sensitivity = 0.90, specificity = 0.88, mean probability for classifying a fall = 0.60). Based on the results of Simon's study, the NQF-endorsed NDNQI patient fall definition was revised to provide more standardized reporting of falls. Although remarkable efforts have been made to improve fall reporting, previous research has indicated a lack of standardized definition and methods of measuring and reporting fall-related injuries (Schwenk et al., 2012). As previously mentioned, injuries associated with falls increase the cost of health care substantially. Without standardized clinical guidelines for reporting injury falls, hospitals lack the ability to properly compare themselves with reliable national comparison data and to develop and implement cost-effective fall prevention plans. Given the financial impact on both hospitals and patients, correct classification of fall-related injuries is imperative, particularly being able to distinguish no injury and minor injuries from serious injuries. Correct classification will allow hospital fall prevention efforts to better target education, risk assessment, and prevention protocols. Thus, the need to

evaluate standardized reporting of injury levels, the key to a reliable and valid injury falls measure, is apparent.

4.1.2 Purpose

The purpose of the study was to investigate the reliability and validity of the NDNQI falls with injury measure by utilizing the NQF and NDNQI injury level definitions (NDNQI, 2010). The specific aims were to assess (a) the consistency of injury level assignment among raters of the fall injury scenarios, and (b) the accuracy of correct injury level assignment. The information on the fall scenarios emulated those commonly found in adverse event or incident reports. Before the study began, approval was obtained from a Midwestern academic medical center Human Subjects Committee (HSC).

4.2 Method

4.2.1 Design

Data collection for the injury falls reliability study followed a similar process to regular falls reporting to NDNQI by member hospitals. When a patient fall occurred in a hospital, a detailed incident report regarding the fall would be filed, including the hospital location of the fall; whether the fall was witnessed, self-reported, or assisted; medication administered to the patient; and any injuries observed at the time of the fall or during post-assessment. Based on the information collected on the incident reports, the fall prevention team would review the incident and determine whether it constituted a unit fall or not, and assign the proper injury level according to NDNQI definitions, which are described in a later section. A unit fall indicates that the event was a fall that occurred on a unit declared eligible by NDNQI for falls reporting. Once

the incident had been thoroughly reviewed, it would be reported to NDNQI along with any other fall incidents on the same unit for the calculation of a unit fall rate.

4.2.2 Participants

Each NDNQI member hospital identifies a site coordinator whose primary responsibility is being a point of contact for all NDNQI-related activities. The NDNQI site coordinator serves a vital role in ensuring that all data collection and reporting adhere to NDNQI guidelines. Thus, the targeted survey population consisted of a convenience sample of site coordinators.

In total, 1,159 site coordinators were invited to participate and 461 responded, resulting in a 39.7% response rate. Among all respondents, 411 provided responses for all fall scenarios that were considered as "complete" responses. Specific instructions for the site coordinators were provided in an email invitation. Because fall prevention programs in hospitals are often viewed as an inter-professional team effort, other hospital staff who serve as final decision makers about injury levels also were asked to be consulted while completing the survey. The most important aspect of the survey was that respondents must assign each scenario to a fall injury level using the NDNQI definitions. A typical respondent was a registered nurse (RN; 91%), held a masters or higher degree (60%), and worked in nursing management (40%) or quality improvement (31%).

4.2.3 Survey Development

A Fall Injury Level Survey was generated using a convenience sample of de-identified incident reports from NDNQI hospitals and NDNQI guidelines on injury levels. Each scenario went through rigorous revisions after being reviewed by hospital and NDNQI staff members who were involved in patient fall-related activities. This process was critical to ensure the content

validity of the fall scenarios on the survey. Twenty fall scenarios were selected as candidates for the final survey.

Two senior NDNQI staff members served as fall experts for determining the correct classification of injury levels in the 20 fall scenarios. Both experts were masters prepared RNs with over 30 years of clinical experience and who provided daily guidance for NDNQI hospitals on classifying actual falls. The experts scored the fall scenarios independently and reached 100% agreement on classification after discussions. Five scenarios were excluded from this study as they were identified by the experts as not a fall or not a unit fall according to the NDNQI fall definition. Thus, the NDNQI experts' judgment was considered the correct injury level classification and deemed to be the "gold" standard. The final Injury Fall Level Survey consisted of 15 fall scenarios, and the distributions of the scenarios were as follows: six non-injurious falls, three minor injury falls, three moderate injury falls, three major injury falls, and zero death resulting from a fall (Table 4.1). Having the experts' gold standard was a crucial first step for subsequent statistical analysis. Table 4.1 shows an abbreviated description and the expert classification for each of the scenarios.

To address the first aim, survey participants were asked to classify the injury level of each scenario according to the NDNQI definitions. Also, questions were included in the survey about the respondents' characteristics such as professional background, highest education level, and current work department within the hospital. The Fall Injury Level Survey was conducted online using the survey tool Zoomerang (http://www.zoomerang.com).

Table 4.1. Expert injury level classification and mean scale score of fall scenarios.

Fall Scenario		Expert Classification	Mean Scale Score [95% CI]
S1[a]	Pt. found sitting on bathroom floor. Steri-strips applied to lacerations on elbow.	Moderate	2.72 [2.26, 3.17]
S2[a]	Pt. lost balance and fell backward. Complained of low back pain. MD ordered Dilaudid and heat packs applied. X-rays negative for fracture or displacement.	Minor	2.02 [1.50, 2.55]
S3[a]	Pt. was found on floor lying next to bed after a loud sound heard from room. No signs/symptoms of injury at that time and at 24 hr post event.	None	1.21 [0.63, 1.79]
S4	Pt. reported to nurse that she "hurt her arm" during fall when walking to BR. No signs of injury and had full ROM. Tylenol administered.	None	1.59 [1.09, 2.09]
S5[a]	Pt. stated he tripped on IV pump power cord and fell. No pain or other injury at the time of the fall or 24 hr post fall.	None	1.05 [0.82, 1.27]
S6[a]	Pt. reported she fell out of a chair to floor while reaching for a book on bedside table. Her NG tube was pulled out, but no other pain or signs of injury 24 hr post fall. MD said to leave NG tube out.	None	1.10 [0.77, 1.42]
S7[a]	Pt. states she fell on knees while reaching for shoes. No injury noted at the time. The next day (15 hr later) pt. complained of R knee pain. X-ray negative, ice, and ACE bandage applied.	Minor	2.04 [1.68, 2.39]
S8[a]	Pt. found on floor. Complained of pain on R side of head, R elbow, and knees. Pt. states he is dizzy, neuro checks found reduced R hand grasp. Small subdural hematoma found on CT scan and pt. transferred to ICU.	Major	3.91 [3.58, 4.25]
S9[a]	Pt. reported he tripped with walker on door jam and fell. Pt. denies pain or other symptoms. Chest X-rays prior to fall indicated a recent rib fracture. Pain meds given 4 hr prior to deep breathing exercises.	None	1.54 [0.54, 2.54]
S10[a]	Pt. found on BR floor and states she hit head. Small laceration on forehead and bandaid applied. Also complained of low back pain, CT of head and lumbar back negative for fracture or hematomas. Pt. given acetaminophen.	Minor	2.09 [1.73, 2.46]
S11[b]	Pt. found unconscious on BR floor after a loud sound heard from room. Large amount of blood on BR floor, sink, and R side of head. Does not respond to painful stimuli, pupils dilated, no B/P, weak and thready pulse. Code blue activated and	Moderate	4.88 [4.32, 5.43]

	CPR performed for 15 min without success.		
S12[a]	While pt. was assisted to BR with gait belt he became dizzy. While trying to lower pt. to the toilet, he became limp and was lowered to the floor. He arm struck the handrail and started swelling. X-ray revealed closed fracture of ulna and a cast was applied.	Major	3.75 [3.27, 4.23]
S13[a]	Pt. walked unassisted to BR after returned to room from EGD. Pt. states he fell to floor after trying to get back in bed. He complained of pain in R ankle. X-ray revealed distal fracture and a cast was applied. After 3 days, pt. complained of numbness and tingling in foot and toes appear blue/purple with swelling. Cast removed 17 hr later by MD and no pedal pulses. Pt. taken to OR for immediate amputation.	Major	3.97 [3.78, 4.16]
S14[b]	Pt. lost balance and fell to floor during transfer from commode to bed. Six staff helped lift pt. with bath blankets to bed and blankets ripped and pt. fell against side rails. Pt. treated for 5 inch abrasion to lumbar area. X-ray of lumbar revealed small compression fracture and treated with back brace.	None	3.60 [2.99, 4.21]
S15[a]	Pt. became dizzy while walking to BR with assistance. Nurse assisted patient to the floor. Pt. sustained 4 inch skin tear on R forearm during the decent. Steri-strips and Kerlix bandage applied.	Moderate	2.66 [2.18, 3.15]

Note. Injury level scale: 1 = *none*, 2 = *minor*, 3 = *moderate*, 4 = *major*, 5 = *death*. CI = confidence interval; Pt. = patient; MD = medical doctor; BR = bathroom; ROM = range of motion; IV = intravenous therapy; NG = nasogastric; R= right; ACE = all cotton elastic (a bandage brand name); CT = computerized tomography; ICU = intensive care unit; B/P = blood pressure; CPR = cardiopulmonary resuscitation; EGD = esophagogastroduodenoscopy; OR = operating room; CFA = confirmatory factor analysis.
a. Final scenario selected by CFA.
b. Complex scenario.

4.2.4 NDNQI Fall and Injury Level Definitions

The NQF-endorsed NDNQI fall and injury level definitions were given in the survey to assist respondents with injury level classifications for the fall scenarios (NDNQI, 2010). A fall was defined as

> an unplanned descent to the floor (or extension of the floor, e.g., trash can or other equipment) with or without injury to the patient, and occurs on an eligible reporting

> nursing unit. All types of falls are to be included whether they result from physiological reasons (fainting) or environmental reasons (slippery floor). Include assisted falls—when a staff member attempts to minimize the impact of the fall. Exclude falls by visitors, students, and staff members; falls on other units not eligible for reporting; falls of patients from eligible reporting units, however patient was not on unit at time of the fall (e.g., patient falls in radiology department). (p. 13)
>
> Injury levels are reported to NDNQI (2010) based on the following guidelines:
>
> None—patient had no injuries (no signs or symptoms) resulting from the fall, if an x-ray, CT scan or other post fall evaluation results in a finding of no injury
> Minor—resulted in application of a dressing, ice, cleaning of a wound, limb elevation, topical medication, pain, bruise or abrasion
>
> Moderate—resulted in suturing, application of steri-strips/skin glue, splinting or muscle/joint strain
>
> Major—resulted in surgery, casting, traction, required consultation for neurological (basilar skull fracture, small subdural hematoma) or internal injury (rib fracture, small liver laceration) or patients with coagulopathy who receive blood products as a result of a fall
>
> Death—the patient died as a result of injuries sustained from the fall (not from physiologic events causing the fall). (pp. 14-15)

4.2.5 Analysis

Coding of responses. Each respondent selected one out of five injury levels according to NDNQI definitions for each of the 15 fall scenarios described in the survey. The response options were coded as 1 “none,” 2 “minor,” 3 “moderate,” 4 “major,” and 5 “death.” The correct injury level for each scenario was the gold standard set by the experts’ classification as described above. Based on the gold standard, all participant responses were further classified as 1 “correct” and 0 “incorrect,” for all 15 fall scenarios. The data file containing the recoded dichotomous data for the 15 fall scenarios served as the main file for all statistical analyses used in this study.

Reliability and validity analysis. The reliability of a measure is the “ability to produce similar results when repeated measurements are made under identical conditions” (Bordens &

Abbott, 2011, p. 130). One common practice to assess the reliability of a target, under the influence of judgments made by a group of respondents, is to calculate the intra-class correlation coefficient (ICC). ICC is calculated as the proportion of the total variance that is due to the true variance from raters (Skrondal & Rabe-Hesketh, 2004). For this study, the fall scenarios were treated as targets and the survey participants as raters. An overall ICC could be used to describe the between-scenario variation of injury level assignment. A high ICC would indicate that the majority of the variance was due to differences among the scenarios, which implied that the difference within each scenario, influenced by raters, was small. Thus, the raters had a high consistency of injury fall classification for each scenario. In this study, the overall ICC estimate was interpreted as excellent (around .90), very good (around .80), and adequate (around .70), following general guidelines provided by Kline (2011). The overall reliability estimate computation was performed using SPSS software version 20.

In addition to reliability, the validity of the fall scenarios also was assessed. The validity of a measure is defined as "the extent to which it measures what you intend it to measure" (Bordens & Abbott, 2011, p. 133). For the 15 fall scenarios, it was important to assess the construct validity of the scenarios. In other words, the goal was to determine if the fall scenarios appropriately could predict the severity of injury falls by assessing the accuracy of correct injury level assignment. A decision needed to be made after examining the proportion of respondents selecting the exactly correct injury level and selecting the correct injury level within one response option, both with a 95% confidence interval. Two scenarios (S11 and S14) were very complex and might have caused a large proportion of the respondents to choose the wrong injury level (Table 4.2). Given the psychometric difficulties, a decision was made to eliminate these two scenarios from the construct validity analysis.

Table 4.2. 95% Confidence interval for the proportion of exactly correct and correct within one injury level.

Fall Scenario[a]	Exactly Correct (%)	Correct Within One Injury Level (%)
S1	[67.17, 75.44]	[100.00, 100.00]
S2	[69.97, 78.04]	[98.59, 100.08]
S3	[82.47, 88.95]	[98.94, 100.17]
S4	[36.52, 45.63]	[98.94, 100.17]
S5	[93.62, 97.45]	[99.34, 100.21]
S6	[88.66, 93.89]	[98.23, 99.98]
S7	[84.09, 90.30]	[100.00, 100.00]
S8	[90.11, 95.00]	[97.89, 99.86]
S9	[69.56, 77.81]	[78.56, 85.74]
S10	[84.27, 90.50]	[98.91, 100.18]
S11[b]	[0.00, 0.68]	[3.39, 7.70]
S12	[73.30, 81.18]	[96.90, 99.42]
S13	[94.53, 98.08]	[100.00, 100.00]
S14[b]	[0.43, 2.82]	[1.75, 5.21]
S15	[61.83, 70.73]	[98.90, 100.18]

a. Abbreviated descriptions of the scenarios are summarized in Table 4.1.
b. Complex scenario.

Thirteen fall scenarios remained for assessment of construct validity which was approached with a two-stage factor analysis using only complete responses. An exploratory factor analysis (EFA) was the logical first step to explore the possible latent factor structure of the injury levels among the fall scenarios. Once the latent factor structure was identified from EFA, it was necessary to verify the factor structure by using a confirmatory factor analysis (CFA) with structural equation modeling. Factor analysis is a correlation-oriented approach that aims to reproduce the inter-correlation among the variables. Several types of correlations exist; however, due to the nature of dichotomous data in this study, tetrachoric correlation was the most appropriate correlational method to serve as the basis of the factor analysis. Unlike Pearson's correlation for continuous data, using tetrachoric correlation allowed us to estimate correlations among dichotomously measured variables as if the variables were made on a continuous scale.

The construct validity computations were performed using Mplus software version 5.21 (L. K. Muthén & Muthén, 1998-2012). Mplus is an advanced statistical software recognized for its powerful ability to fit various latent variable models. Following recommendations by MacCallum, Roznowski, and Necowitz (1992), the main analysis file with 411 complete responses were randomly split into comparable training (196 responses, 47.7%) and validation (215 responses, 52.3%) data sets to avoid capitalization on chance concerns. An EFA with categorical factor indicators was conducted using the training data set in Mplus, which conveniently incorporated tetrachoric correlation into the analysis. Traditional factor extraction, such as Kaiser's criterion, has been accepted widely for suggesting factors with an eigenvalue greater than one as common factors. Eigenvalues often are interpreted as the variances extracted by the common factors. However, eigenvalues based Kaiser's criterion should not be used solely to determine the number of factors due to over-extraction concerns. Another requirement for including items in a specific factor was that the individual items must meet a criterion of at least 0.30 in absolute value for factor loading to be retained. Additional model fit can be evaluated by using the root mean square error of approximation (RMSEA), and a RMSEA value around 0.05 or less usually indicates an acceptable model fit. As latent factors were identified, a CFA with categorical factor indicators using structural equation modeling was performed on the validation data set to confirm the factor structure demonstrated in the EFA step. Several statistical indices such as the comparative fit index (CFI; around 0.9 or higher), Tucker Lewis Index (TLI; around 0.9 or higher) and RMSEA (around 0.05 or less) were used to assess the final model fit.

For oblique rotations (correlated factors, for example, Promax), the concept of the proportion of variance explained by a factor is complex and less intuitive. Factor solutions provided by a Varimax rotation (uncorrelated factors) are often very similar to the Promax solutions. Thus, the

Varimax factor solutions can be used as a proxy to compute the variability explained by a given factor under the Promax setting. The proportion of variance explained by a factor can be calculated as the sum of squared factor loadings on the assigned factor divided by the number of fall scenarios assigned to that particular factor. In addition, Mplus also provides estimates for the proportion of variance in each fall scenario, explained by their assigned factor.

4.3 Results

4.3.1 Reliability

The variance within each scenario was 0.252 and the variance between the 15 fall scenarios was 1.479, resulting in an overall ICC (1, 1) of .85, which was between "very good" and "excellent" according to the general guidelines provided by Kline (2011). The ICC (1, 1) indicated a substantial reliability of the fall scenarios and a high consistency of injury level assignment among the respondents for each scenario. The mean scale scores with 95% confidence intervals for all 15 fall scenarios are summarized in Table 4.1 (p. 67). The variance between the scenarios was much larger than the variance within each scenario, which echoed the results of the overall ICC estimate and indicated a high reliability.

As mentioned above, two scenarios (S11 and S14) were very complex and were excluded from further analysis. After exclusion, the overall ICC (1, 1) for the remaining 13 scenarios was re-calculated to be .82, which still maintained a very good reliability and was suitable for the validity analysis.

4.3.2 Validity

During the initial EFA conducted on the training data set, six factors with eigenvalues greater than 1 were suggested based on Kaiser's criterion (eigenvalues: 3.556, 2.807, 1.582, 1.195, 1.171, 1.079, 0.692, 0.654, 0.353, 0.284, 0.094, –0.163, –0.305), but only three factors could be extracted successfully, indicating an over-extraction based on Kaiser's criterion. Factor loadings of the three-factor model were further clarified after applying a Promax rotation for correlated factors, resulting in a RMSEA of 0.053, which indicated an acceptable model fit. All scenarios loaded over 0.30 on the assigned factors. The aim of the EFA was to identify underlying factor structure that could be used to predict the severity of injury falls. The results indicated three latent factors: ability associated with classifying non-injurious falls (No Injury), ability associated with classifying minor injury falls (Minor Injury), and ability associated with classifying moderate or major injury falls (Moderate/Major Injuries; Table 4.3).

Table 4.3. Factor Loadings After Promax Rotation for Three-Factor Structure With Injury Levels.

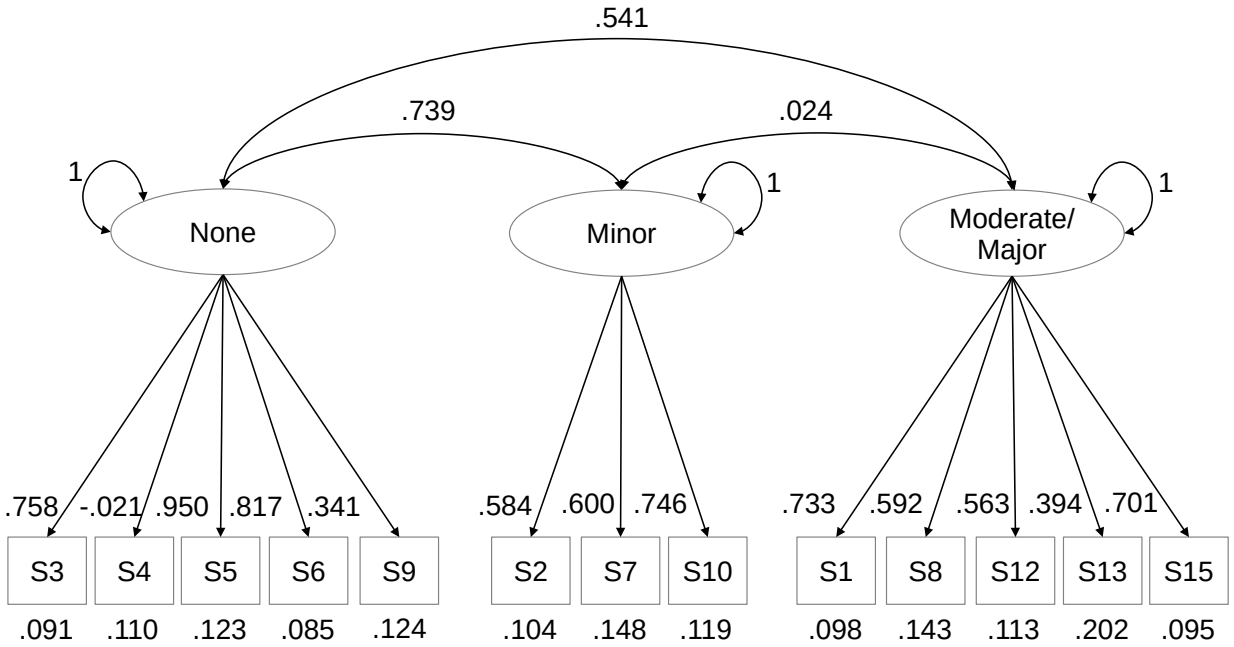
Fall Scenario[a]	No Injury	Minor Injury	Moderate/Major Injuries	Injury Level
S1	–0.078	0.239	**0.801**	Moderate
S8	–0.173	–0.014	**0.535**	Major
S12	0.104	–0.005	**0.643**	Major
S13	0.005	–0.144	**0.883**	Major
S15	0.094	0.028	**0.715**	Moderate
S2	–0.019	**0.778**	0.033	Minor
S7	0.059	**0.504**	–0.005	Minor
S10	0.197	**0.312**	–0.274	Minor
S3	**0.448**	0.444	0.233	None
S4	**0.758**	–0.244	0.008	None
S5	**0.873**	0.64	–0.023	None
S6	**0.684**	0.393	–0.064	None
S9	**0.311**	0.03	0.082	None

a. Abbreviated descriptions of the scenarios are summarized in Table 4.1.
Note. Highest factor loading for each fall scenario is in bold.

With the validation data set, the CFA model was specified using the three factors measured by the 13 scenarios, with each scenario assigned to the relevant factor. The goal was to identify and retain scenarios that contributed most to respondents' ability associated with injury fall classifications. Estimates of the pattern coefficients representing the direct effects of the factors on the scenarios ranged from –0.021 to 0.950 (Figure 4.1A). Several statistical indices were used to determine the adequacy of model fit to the data. Results from the initial CFA assessment did not indicate a good model fit (CFI = 0.868, TLI = 0.863, RMSEA = 0.055). Pattern coefficient estimates for all scenarios were statistically significant (p value < .05) with the exception of Scenario 4 (–0.021, p value = .851) and Scenario 13 (0.395, p value = .051). The pattern coefficient estimate for Scenario 13 can be considered as marginally significant and we decided to keep this scenario in the model. The CFA model was refitted after removing Scenario 4 and the confirmed structure remained the same (Figure 4.1B). The final CFA assessment confirmed an acceptable model fit and supported the hypothesis that a relationship exists between the 12 final fall scenarios (Table 4.1, p. 67) and the three underlying latent factors (CFI = 0.914, TLI = 0.910, RMSEA = 0.048).

As mentioned above, Varimax factor solutions were used as a proxy to calculate the variability explained by the three correlated latent factors. Results from the Varimax rotation are not reported here due to the high degree of similarity with the Promax rotation solutions. The proportion of variance explained by the No Injury, Minor Injury, and Moderate/Major Injuries factors were 52.3%, 31.9%, and 46.7%, respectively. In addition, the ability associated with classifying non-injurious falls accounted for 53.6%, 34.8%, 31.8%, 15.6%, and 49.3% of the proportion of variance in Scenarios 1, 8, 12, 13, and 15, respectively. The variability in Scenarios 2, 7, and 10, explained by the ability associated with classifying minor injury falls, were 33.7%,

A



B

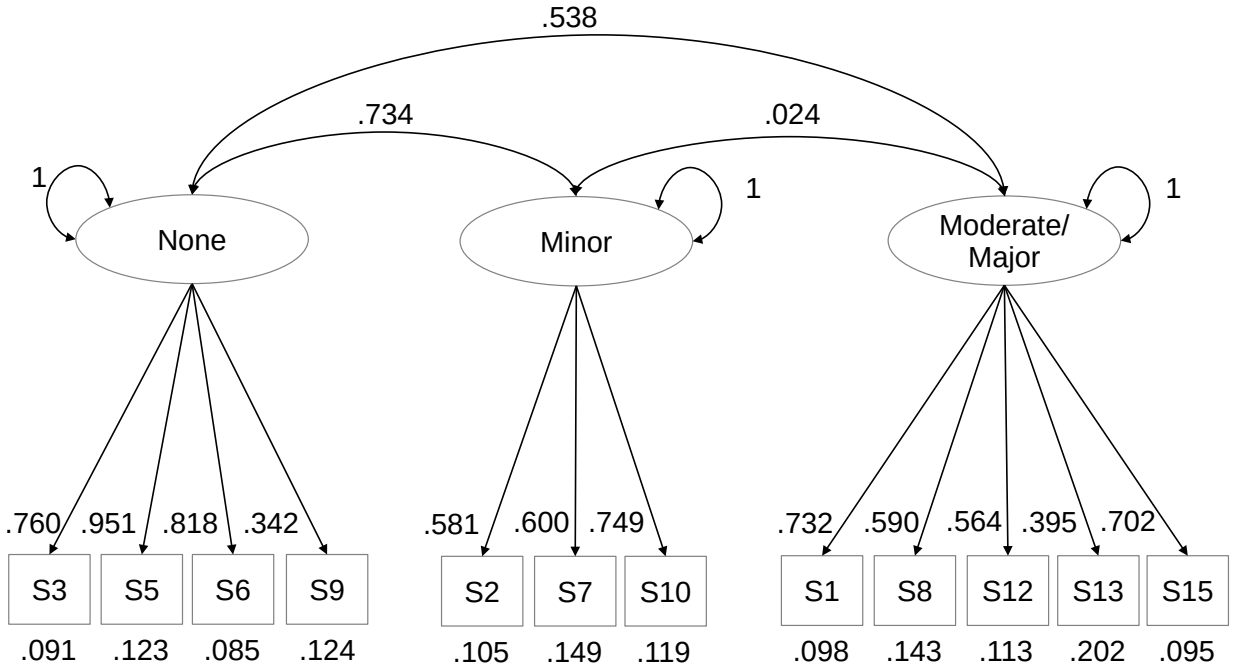


Figure 4.1. Initial CFA model (A) and final CFA model (B).
Note. CFA = confirmatory factor analysis.

36.0%, and 56.1%, respectively. Finally, the ability associated with classifying moderate or major injury falls accounted for 57.8%, 90.5%, 66.9%, and 11.7% of the variability in Scenarios 3, 5, 6, and 9, respectively.

The construct validity analysis findings indicated that the final 12 fall scenarios from the survey resulted in appropriate latent structures for predicting the severity of the injury falls, and thus supporting the validity or accuracy of injury level classifications made by survey respondents for all 12 final fall scenarios.

4.4 Discussion

The overall ICC estimate for the 15 fall scenarios fell between very good and excellent, indicating high consistency of injury level classifications among respondents for each fall scenario. Results provided strong evidence for the reliability of the NDNQI falls with injury measure. Construct validity also was confirmed, resulting in 12 final fall scenarios with four non-injurious falls, three minor injury falls, two moderate injury falls, and three major injury falls. The 12 final fall scenarios represented a reliable and valid approach to evaluate respondent fall injury level classification ability.

From the results of the construct validity analysis, it was apparent that the scenarios clustered very well into the three distinct categories. However, the correlations among the three latent factors exhibited a very interesting pattern, that could be presented as poor (Minor vs. Moderate/Major = .024, *p* value = .810), average (None vs. Moderate/Major = .538, *p* value < .05), and good (None vs. Minor = .734, *p* value < .05). The pattern in the factor correlation estimates merited further investigation. The poor correlation (.024) between Minor Injury and Moderate/Major Injuries could be interpreted, such that the respondents' ability to correctly classify minor injuries did not imply that they also would have the same ability to correctly

classify moderate or major injuries, and vice versa. This finding is rather concerning and can indicate several potential issues, such as confusion over the definitions, ambiguity of the incident reports, or bias introduced from both the patient and fall evaluator's perspectives. On the contrary, it is certainly encouraging to see that the respondents had average ability to correctly distinguish no injury from moderate or major injuries, and vice versa. Moreover, the respondents had a good ability to correctly classify no injury from minor injuries, and vice versa. The overall results can be viewed as an indication that more education or training is needed for correctly identifying all injury levels, particularly the moderate or major injury falls, as these types of fall scenarios are rare. The clarity of the injury level definitions also needs to be further reviewed to minimize potential classification challenges. In addition, although the construct validity assessed respondents' ability to distinguish among No Injury, Minor Injury, and Moderate/Major Injuries, the ability to distinguish injury levels within the global category of Moderate/Major Injuries remains unknown and requires further investigation.

The majority of fall scenarios had about 70% to 90% of respondents selecting the exactly correct injury level with the exception of three scenarios (S4, S11, and S14). Specifically S11 and S14 had close to 0% of the respondents being exactly correct (Table 4.2, p. 71). When the requirement was relaxed to allow within one injury level, S11 and S14 still remained very low with less than 10% of the respondents being correct (Table 4.2). The sequence of events in Scenario 11 made it unclear whether the fall caused the patient death or the death caused the fall. In Scenario 14, the patient fell and then was dropped by the staff as they attempted to assist the patient back to bed, leading to confusion about the injury level assignment. These two scenarios were considered to be very complex, which resulted in a wide variance of injury level assignment among the respondents. Thus, both fall scenarios were excluded from the construct

validity analysis for psychometric difficulties. The complex fall scenarios (e.g., S11 and S14) need to be examined carefully and debriefed by the fall prevention team, and when necessary, expert consultations should be considered to help prevent bias by the fall evaluator. In addition, concerns can arise with patient self-reported falls (e.g., S4) because this type of fall often is not observed and hard to validate without evidence; thus, potential bias could be introduced from both the patient and fall evaluator's perspectives.

One limitation of this study comes from the usage of incident reports to help design the online survey. Previous research by Shorr and colleagues (2008) pointed out that using incident reports alone contributes to the underreporting of both injurious and non-injurious falls in hospitals. Potential bias could be introduced by using a convenience sample of de-identified incident reports that are not representative for all fall scenarios that patients experience daily in hospitals. Although all fall scenarios went through rigorous revisions to ensure their clinical reality, it remains unclear how frequent these scenarios occur. Perhaps more scenarios need to be developed to cover the full spectrum of NDNQI injury classifications.

Another limitation of this study comes from the sample selection bias. The primary audience for the survey was a convenience sample of NDNQI site coordinators. Comparing with the general population of U.S. hospitals, NDNQI consists of more Magnet®-designated, not-for-profit, larger, and higher case-mix index (CMI) hospitals (Lake, Shang, Klaus, & Dunton, 2010). The general profile of NDNQI hospitals may include more hospital resources that play an important role in establishing training for staff and fall prevention programs. Being the primary respondent of the survey (68%), NDNQI site coordinators are constantly informed on new updates to NDNQI guidelines and definitions. They are most familiar with NDNQI frameworks and thus may represent a more "trained" group of hospital staff in regard to standardized data

collection and reporting. The ability of correct injury level classification across other hospital staff involved in fall related activities still remains unclear and needs to be further evaluated.

In this study, the reliability and validity of the NDNQI falls with injury measure was evaluated and findings supported the successful re-endorsement by NQF. The NDNQI site coordinators demonstrated high consistency in classifying injury levels for specific fall scenarios, according to NDNQI definitions. The Falls Injury Level Survey with the final 12 fall scenarios was shown to be valid in assessing respondents' abilities to predict the severity of the injury falls, particularly among non-injurious falls, minor injury falls, and moderate or major injury falls. Hospital site coordinators are encouraged to continue contacting NDNQI for assistance with the classification of complex fall scenarios and patient self-reported fall scenarios. Findings of this study also supported rationales for revising the standardized NDNQI falls and injury level definitions to include additional types of falls and provide more clarification on injuries.

An implication from this study is that the Falls Injury Level Survey can be utilized in the future as a training tool for hospital staff that serve as final decision makers on injury levels. Researchers at NDNQI launched a well-known and comprehensive Pressure Ulcer Identification and Staging Training Program in 2009 that can be used to guide the development of a falls with injury training tool (Bergquist-Beringer et al., 2009; Bergquist-Beringer, Gajewski, Dunton, & Klaus, 2011; Gajewski, Hart, Bergquist-Beringer, & Dunton, 2007; Hart, Bergquist, Gajewski, & Dunton, 2006). In addition, because the NDNQI injury falls measure is NQF-endorsed, standardized injury level definitions are available to the public domain. A recent article published by Mion and colleagues (2012) utilized NDNQI injury level definitions as part of their retrospective study for determining potential predictors and outcomes of injurious falls among a cohort of hospital patients. The NDNQI injury falls measure provides a reliable and valid tool for

non-NDNQI hospitals and external researchers to support future quality improvement efforts and injurious falls research.

Chapter Five

Summary and Future Directions

The role of biostatisticians in solving health care-related issues have become increasingly important, especially for today's U.S. health care where patient-centeredness is recognized as a national priority. To promote quality of care, it is essential for clinicians, health care researchers, and biostatisticians to work in close collaboration for the development of reliable and valid PROMs and ClinRO measures. While clinicians and health care researchers contribute significantly in identifying and developing concepts of interest, biostatisticians offer valuable expertise in proposing novel statistical methods beyond the traditional approaches. This dissertation provides a thorough overview of a novel Bayesian method for expediting the development of PROMs and an application of traditional (i.e., frequentist) instrument development methods in the psychometric evaluation of a ClinRO measure.

Traditional psychometric methodologies are efficient and reliable when developing PROMs for populations with relatively large sample sizes. However, in practice, researchers may not always have access to a large participant pool (i.e., in cases of rare diseases) that is required for classical psychometric assessments. Additional challenges such as a lengthy process and/or limited resources can further cripple a classical instrument development process. An innovative Ordinal Bayesian Instrument Development (OBID) approach within a Bayesian IRT framework is proposed in this dissertation to overcome both small sample size and ordinal data modeling challenges (Manuscript 1 in Chapter Two). Subject experts' opinions (content validity) are incorporated seamlessly and efficiently under the OBID approach to form the prior distributions for the IRT parameters in the participant data model (construct validity). The efficiency of OBID is evaluated by comparing its performance to classical instrument development performance under a simulation setting with three different types of expert bias. Results successfully demonstrated the superior performance of the OBID approach with small sample sizes; thus

OBID offers a reliable alternative for future PROMs development for small populations or rare diseases.

As previously mentioned, the proposed OBID approach is developed using a Bayesian IRT model framework. Literature has indicated that the assessment of IRT model fits is both a challenging and an underdeveloped area in research (Sinharay & Johnson, 2003; Sinharay et al., 2006). In this dissertation, an exact Bayesian LOO-CV approach is investigated to compare Bayesian IRT models in PROMs development (Manuscript 2 in Chapter Three). Results support the incorporation of appropriate content subject experts' information in establishing construct validity under the OBID approach. However, the appropriate selection of subject experts is an important area to focus in order to efficiently implement the OBID approach and reduce potential bias during PROMs development.

Despite increasing public awareness on the concept of patient-centered care, the development of ClinRO measures is equally important in promoting the quality of health care. Existing ClinRO measures require routine re-assessment to maintain the psychometric integrity of the measure. This is especially critical for measures that are endorsed by national entities, as these measures provide standardization and comparability for hospitals' quality improvement efforts. The reliability and validity of the NQF-endorsed NDNQI falls with injury measure was evaluated in this dissertation (Manuscript 3 in Chapter Four). Findings of the study not only supported the successful re-endorsement of the injury falls measure by NQF, the final Falls Injury Level Survey also was shown to be valid in assessing respondents' abilities to predict the severity of the injury falls. The NDNQI injury falls measure is a reliable and valid ClinRO measure for future quality improvement efforts and injurious falls research.

This dissertation has motivated several topics that can be considered for future studies. First, the application capability of OBID can be extended through the development of a user-friendly software called Classical & Bayesian Instrument Development (CBID; Karanevich et al., in review). Second, a hierarchical model can be considered to incorporate the individual effect of content experts, as the scores experts assigned from item to item are likely to be correlated. Third, an approximation to the Bayesian LOO-CV approach for ordinal latent variable models can be explored to improve the efficiency of Bayesian cross-validation in IRT models. Fourth, OBID subject experts' model further can be evaluated through simulation with more skewed participant data structure and other prior distributions. Last but not least, motivated from the ClinRO measure evaluation study, a falls with injury training tool can be developed in the future for hospital staff that serve as final decision makers on fall injury levels.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Proceedings of the Second International Symposium on Information Theory* (pp. 267-281). Budapest: Akademiai Kiado.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Anthoine, E., Moret, L., Regnault, A., Sébille, V., & Hardouin, J. B. (2014). Sample size used to validate a scale: a review of publications on newly-developed patient reported outcomes measures. *Health and quality of life outcomes, 12*, 176.

Arima, S. (2015). Item selection via Bayesian IRT models. *Stat Med, 34*(3), 487-503. doi:10.1002/sim.6341

Basch, E., Bennett, A., & Pietanza, M. C. (2011). Use of Patient-Reported Outcomes to Improve the Predictive Accuracy of Clinician-Reported Adverse Events. *Journal of the National Cancer Institute, 103*(24), 1808-1810. doi:10.1093/jnci/djr493

Bergquist-Beringer, S., Davidson, J., Agosto, C., Linde, N. K., Abel, M., Spurling, K., . . . Christopher, A. (2009). Evaluation of the National Database of Nursing Quality Indicators (NDNQI) Training Program on Pressure Ulcers. *J Contin Educ Nurs, 40*(6), 252-258; quiz 259-260, 279. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19639914

Bergquist-Beringer, S., Gajewski, B., Dunton, N., & Klaus, S. (2011). The reliability of the National Database of Nursing Quality Indicators pressure ulcer indicator: a triangulation approach. *J Nurs Care Qual, 26*(4), 292-301. doi:10.1097/NCQ.0b013e3182169452

Bordens, K., & Abbott, B. B. (2011). *Research Design and Methods A Process Approach: Eighth Edition*: McGraw-Hill Higher Education.

Breiman, L., & Spector, P. (1992). Submodel Selection and Evaluation in Regression - the X-Random Case. *International Statistical Review, 60*(3), 291-319. doi:Doi 10.2307/1403680

Brown, T. A. (2014). *Confirmatory factor analysis for applied research* (2nd ed.): Guilford Publications.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., . . . Group, P. C. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *J Clin Epidemiol, 63*(11), 1179-1194. doi:10.1016/j.jclinepi.2010.04.011

Centers for Disease Control and Prevention, National Center for Injury Prevention and Control, & Division of Unintentional Injury Prevention. (2013). *Costs of Falls Among Older Adults*. Retrieved from http://www.cdc.gov/homeandrecreationalsafety/falls/fallcost.html.

Centers for Medicare & Medicaid Services. (2012). *Hospital-Acquired Conditions (HAC) in Acute Inpatient Prospective Payment System (IPPS) Hospitals*. Baltimore, MD: Author. Retrieved from http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/HospitalAcqCond/Downloads/HACFactsheet.pdf.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cowles, M. K. (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing, 6*(2), 101-111. doi:Doi 10.1007/Bf00162520

Cuijpers, P., Li, J. A., Hofmann, S. G., & Andersson, G. (2010). Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: A meta-analysis. *Clinical Psychology Review, 30*(6), 768-778. doi:10.1016/j.cpr.2010.06.001

Currie, L. (2008). Fall and Injury Prevention. In R. G. Hughes (Ed.), *Patient Safety and Quality: An Evidence-Based Handbook for Nurses*. Rockville (MD).

Davidian, M., & Louis, T. A. (2012). Why statistics? *Science, 336*(6077), 12. doi:10.1126/science.1218685

Dawson, J., Doll, H., Fitzpatrick, R., Jenkinson, C., & Carr, A. J. (2010). The routine use of patient reported outcome measures in healthcare settings. *BMJ, 340*, c186. doi:10.1136/bmj.c186

Deisseroth, A., Kaminskas, E., Grillo, J., Chen, W., Saber, H., Lu, H. L., . . . Pazdur, R. (2012). U.S. Food and Drug Administration Approval: Ruxolitinib for the Treatment of Patients with Intermediate and High-Risk Myelofibrosis. *Clinical Cancer Research, 18*(12), 3212-3217. doi:10.1158/1078-0432.Ccr-12-0653

Eaton, W. W., Smith, C., Ybarra, M., Muntaner, C., & Tien, A. (2004). Center for Epidemiologic Studies Depression Scale: review and revision (CESD and CESD-R). In M. E. Maruish (Ed.), *The Use of Psychological Testing for Treatment Planning and Outcomes Assessment* (3rd ed., pp. 363-377). Mahwah, NJ: Lawrence Erlbaum.

Engelman, K. K., Daley, C. M., Gajewski, B. J., Ndikum-Moffor, F., Faseru, B., Braiuca, S., . . . Greiner, K. A. (2010). An assessment of American Indian women's mammography experiences. *BMC Womens Health, 10*, 34. doi:10.1186/1472-6874-10-34

Engelman, K. K., Ndikum-Moffor, F. M., Gajewski, B. J., Yu, Q., Nazir, N., Daley, C. M., & Ellerbeck, E. F. (in review). Reliability and Validation of a Patient Assessment of Mammography Services (PAMS) Satisfaction Survey.

Englander, F., Hodson, T. J., & Terregrossa, R. A. (1996). Economic dimensions of slip and fall injuries. *J Forensic Sci, 41*(5), 733-746. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8789837

Fox, J. P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*(2), 271-288. doi:Doi 10.1007/Bf02294839

Gajewski, B. J., Coffland, V., Boyle, D. K., Bott, M., Price, L. R., Leopold, J., & Dunton, N. (2012). Assessing Content Validity Through Correlation and Relevance Tools A Bayesian Randomized Equivalence Experiment. *Methodology-European Journal of Research Methods for the Behavioral and Social Sciences, 8*(3), 81-96. doi:10.1027/1614-2241/A000040

Gajewski, B. J., Hart, S., Bergquist-Beringer, S., & Dunton, N. (2007). Inter-rater reliability of pressure ulcer staging: Ordinal probit Bayesian hierarchical model that allows for uncertain rater response. *Stat Med, 26*(25), 4602-4618. doi:Doi 10.1002/Sim.2877

Gajewski, B. J., Price, L. R., Coffland, V., Boyle, D. K., & Bott, M. J. (2013). Integrated analysis of content and construct validity of psychometric instruments. *Quality & Quantity, 47* 57-78.

Garrard, L., Price, L. R., Bott, M. J., & Gajewski, B. J. (2015). A novel method for expediting the development of patient-reported outcome measures and an evaluation of its performance via simulation. *BMC Medical Research Methodology, 15*(1), 77. doi:10.1186/s12874-015-0071-5

Gelfand, A. E., Dey, D. K., & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian Statistics* (4 ed., pp. 147-167): Oxford University Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis. Texts in statistical science series*.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing, 24*(6), 997-1016. doi:10.1007/s11222-013-9416-2

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*(4), 733-760.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2015). mvtnorm: Multivariate Normal and t Distributions. R package version 1.0-3. Retrieved from http://CRAN.R-project.org/package=mvtnorm

Gershon, R. C., Lai, J. S., Bode, R., Choi, S., Moy, C., Bleck, T., . . . Cella, D. (2012). Neuro-QOL: quality of life item banks for adults with neurological disorders: item development and calibrations based upon clinical and general population testing. *Quality of Life Research, 21*(3), 475-486. doi:10.1007/s11136-011-9958-8

Gershon, R. C., Rothrock, N., Hanrahan, R., Bass, M., & Cella, D. (2010). The use of PROMIS and assessment center to deliver patient-reported outcome measures in clinical research. *Journal of applied measurement, 11*(3), 304.

Gibbs, H., & Chapman-Novakofski, K. (2013). Establishing Content Validity for the Nutrition Literacy Assessment Instrument. *Preventing Chronic Disease, 10*. doi:10.5888/pcd10.120267

Gibbs, H. D., Ellerbeck, E. F., Befort, C., Gajewski, B., Kennett, A. R., Yu, Q., . . . Sullivan, D. K. (2015). Measuring Nutrition Literacy in Breast Cancer Patients: Development of a Novel Instrument. *J Cancer Educ*. doi:10.1007/s13187-015-0851-y

Grant, J. S., & Davis, L. L. (1997). Selection and use of content experts for instrument development. *Research in Nursing & Health, 20*(3), 269-274.

Hart, S., Bergquist, S., Gajewski, B., & Dunton, N. (2006). Reliability testing of the National Database of Nursing Quality Indicators pressure ulcer indicator. *J Nurs Care Qual, 21*(3), 256-265. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16816607

IBM Corp. (2011). IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp.

Inouye, S. K., Brown, C. J., & Tinetti, M. E. (2009). Medicare nonpayment, hospital falls, and unintended consequences. *N Engl J Med, 360*(23), 2390-2393. doi:10.1056/NEJMp0900963

Institute of Medicine. (2001). *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, DC: National Academy Press.

ISPOR Clinical Outcomes Assessment – Emerging Good Practices Task Force. (2015). *Clinical Outcome Assessments (COAs) & Clinician Reported Outcomes (ClinROs)*. Paper

presented at the 20th Annual International Meeting, Philadelphia, PA. https://www.ispor.org/TaskForces/COA-ClinRO-TF-presentation-2015Philadelphia.pdf

James, G., Witten, D., & Hastie, T. (2014). An Introduction to Statistical Learning: With Applications in R: Taylor & Francis.

Jiang, Y., Boyle, D. K., Bott, M. J., Wick, J. A., Yu, Q., & Gajewski, B. J. (2014). Expediting Clinical and Translational Research via Bayesian Instrument Development. *Applied psychological measurement*. doi:10.1177/0146621613517165

Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling*. New York, NY: Springer Science & Business Media.

Karanevich, A. G., Garrard, L., Bott, M., Price, L. R., Mudaranthakam, D. P., & Gajewski, B. (in review). Classical & Bayesian Instrument Development: a Free, Easily-Accessible Confirmatory Factor Analysis Software Alternative.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Knapp, T. R. (1990). Treating ordinal scales as interval scales: an attempt to resolve the controversy. *Nursing research, 39*(121-123).

Knapp, T. R., & Brown, J. K. (1995). Ten measurement commandments that often should be broken. *Research in Nursing & Health, 18*, 465-469.

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection.* Paper presented at the Ijcai.

Lake, E. T., Shang, J., Klaus, S., & Dunton, N. E. (2010). Patient falls: Association with hospital Magnet status and nursing unit staffing. *Res Nurs Health, 33*(5), 413-425. doi:10.1002/nur.20399

Li, L., Qiu, S., Zhang, B., & Feng, C. X. (2014). Approximating cross-validatory predictive evaluation in Bayesian latent variables models with integrated IS and WAIC. *arXiv preprint arXiv:1404.2918*.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10*(4), 325-337. doi:10.1023/A:1008929526011

Lynn, M. R. (1986). Determination and Quantification of Content Validity. *Nursing research, 35*(6), 382-385.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychol Bull, 111*(3), 490-504. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16250105

Marshall, S., Haywood, K., & Fitzpatrick, R. (2006). Impact of patient-reported outcome measures on routine practice: a structured review. *J Eval Clin Pract, 12*(5), 559-568. doi:10.1111/j.1365-2753.2006.00650.x

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software, 42*(9), 1-21.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York, NY: American Council on Education.

Mion, L. C., Chandler, A. M., Waters, T. M., Dietrich, M. S., Kessler, L. A., Miller, S. T., & Shorr, R. I. (2012). Is it possible to identify risks for injurious falls in hospitalized patients? *Joint Commission journal on quality and patient safety/Joint Commission Resources, 38*(9), 408. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3547233/pdf/nihms432790.pdf

Montalvo, I. (2007). The National Database of Nursing Quality IndicatorsTM (NDNQI®). *OJIN: The Online Journal of Issues in Nursing, 12*(3). Retrieved from http://www.nursingworld.org/MainMenuCategories/ANAMarketplace/ANAPeriodicals/OJIN/TableofContents/Volume122007/No3Sept07/NursingQualityIndicators.aspx?%3E

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*, 115-132.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus User's Guide* (Seventh ed.). Los Angeles, CA: Muthén & Muthén.

National Database of Nursing Quality Indicators. (2010). *Guidelines for data collection on the American Nurses Association's National Quality Forum endorsed measures*. Kansas City, KS: University of Kansas Medical Center.

National Priorities Partnership. (2011). *Input to the Secretary of Health and Human Services on Priorities for the National Quality Strategy*. Washington, DC: National Quality Forum. Retrieved from http://www.qualityforum.org/WorkArea/linkit.aspx?ItemID=68238.

National Quality Forum. (2002). *A national framework for healthcare quality measurement and reporting: A consensus report*. Washington, DC: Author. Retrieved from http://www.qualityforum.org/Publications/2002/07/A_National_Framework_for_Healthcare_Quality_Measurement_and_Reporting.aspx.

National Quality Forum. (2004). *National voluntary consensus standards for nursing-sensitive care: An initial performance measure set*. Washington, DC: Author. Retrieved from http://www.qualityforum.org/Publications/2004/10/National_Voluntary_Consensus_Standards_for_Nursing-Sensitive_Care__An_Initial_Performance_Measure_Set.aspx.

National Quality Forum. (2011). *Serious reportable events in healthcare—2011 Update: A consensus report*. Washington, DC: Author. Retrieved from http://www.qualityforum.org/Publications/2011/12/Serious_Reportable_Events_in_Healthcare_2011.aspx.

National Quality Forum. (2013a). *Falls with injury*. Retrieved from http://www.qualityforum.org/QPS/0202.

National Quality Forum. (2013b). *Patient fall rate*. Retrieved from http://www.qualityforum.org/QPS/0141.

National Quality Forum. (2013c). *Patient Reported Outcomes (PROs) in Performance Measurement*. Retrieved from http://www.qualityforum.org/Publications/2012/12/Patient-Reported_Outcomes_in_Performance_Measurement.aspx.

Ndikum-Moffor, F. M., Braiuca, S., Gajewski, B. J., Daley, C. M., Yu, Q., & Engelman, K. K. (in review). Focus groups and content validity indexing utilization in the development of a patient assessment of mammography services instrument for American Indian women.

Nunnally, I. H., & Bernstein, J. C. (1994). *Psychometric theory*: New York McGraw-Hill.

Patient Centered Outcomes Research Institute. (2012). *The Design and Selection of Patient-Reported Outcomes Measures (PROMs) for Use in Patient Centered Outcomes Research*. Retrieved from http://www.pcori.org/assets/The-Design-and-Selection-of-Patient-Reported-Outcomes-Measures-for-Use-in-Patient-Centered-Outcomes-Research1.pdf.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research.*: Sage.

Polit, D. F., & Beck, C. T. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*, 489-497.

Price, L. R. (in press). *Psychometric Methods: Theory into Practice*. New York, NY: Guilford Publications.

Quinn, K. M. (2004). Bayesian factor analysis for mixed ordinal and continuous responses. *Political Analysis, 12*(4), 338-353. doi:Doi 10.1093/Pan/Mph022

R Core Team. (2015). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Radloff, L. S. (1977). The CES-D scale a self-report depression scale for research in the general population. *Applied psychological measurement, 1*(3), 385-401.

Rosner, B. (2010). *Fundamentals of biostatistics*: Cengage Learning.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*, 1-36.

Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *Annals of Statistics, 12*(4), 1151-1172. doi:DOI 10.1214/aos/1176346785

Sanderson, T., & Kirwan, J. (2009). Patient-reported outcomes for arthritis: time to focus on personal life impact measures? *Arthritis Rheum, 61*(1), 1-3. doi:10.1002/art.24270

Schwenk, M., Lauenroth, A., Stock, C., Moreno, R. R., Oster, P., McHugh, G., . . . Hauer, K. (2012). Definitions and methods of measuring and reporting on injurious falls in randomised controlled fall prevention trials: a systematic review. *BMC Med Res Methodol, 12*, 50. doi:10.1186/1471-2288-12-50

Shorr, R. I., Mion, L. C., Chandler, A. M., Rosenblatt, L. C., Lynch, D., & Kessler, L. A. (2008). Improving the capture of fall events in hospitals: combining a service for evaluating inpatient falls with an incident report system. *Journal of the American Geriatrics Society, 56*(4), 701-704.

Simon, M., Klaus, S., Gajewski, B. J., & Dunton, N. (2013). Agreement of fall classifications among staff in US hospitals. *Nursing research, 62*(2), 74-81.

Sinharay, S., & Johnson, M. S. (2003). *Simulation studies applying posterior predictive model checking for assessing fit of the common item response theory models* (ETS RR-03-28). Princeton, NJ: Educational Testing Service.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied psychological measurement, 30*(4), 298-321. doi:10.1177/0146621605285517

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*: CRC Press.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 76*(3), 485-493. doi:10.1111/rssb.12062

Spiegelhalter, D. J., Best, N. G., Carlin, B. R., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society Series B-Statistical Methodology, 64*, 583-616. doi:10.1111/1467-9868.00353

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*: Harvard University Press.

Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 39*, 44–47.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago University of Chicago Press.

US Department of Health and Human Services. (2011). *Healthy People 2020 topics and objectives: cancer*. Retrieved from http://www.healthypeople.gov/2020/topicsobjectives2020/objectiveslist.aspx?topicId=5.

US Department of Health and Human Services Food and Drug Administration. (2009). *Guidance for Industry Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims*. Retrieved from http://www.fda.gov/downloads/Drugs/Guidances/UCM193282.pdf.

US Department of Health and Human Services Food and Drug Administration. (2014). *Guidance for Industry and FDA Staff Qualification Process for Drug Development Tools*. Retrieved from http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm230597.pdf.

US Department of Health and Human Services Food and Drug Administration. (2015a). *Clinical Outcome Assessment (COA): Glossary of Terms*. Retrieved from http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm370262.htm#ObsRO.

US Department of Health and Human Services Food and Drug Administration. (2015b). *Clinical Outcome Assessment Qualification Program*. Retrieved from

http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm284077.htm.

Vehtari, A., & Gelman, A. (2015). Pareto smoothed importance sampling. *arXiv:1507.02646*.

Vehtari, A., Gelman, A., & Gabry, J. (2015). Efficient implementation of leave-one-out cross-validation and WAIC for evaluating fitted Bayesian models. *arXiv preprint arXiv:1507.04544*.

Vehtari, A., & Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation, 14*(10), 2439-2468. doi:10.1162/08997660260293292

Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys, 6*, 142-228.

Vehtari, A., Tolvanen, V., Mononen, T., & Winther, O. (2014). Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *arXiv preprint arXiv:1412.7461*.

Verstovsek, S., Mesa, R. A., Gotlib, J., Levy, R. S., Gupta, V., DiPersio, J. F., . . . Kantarjian, H. M. (2012). A double-blind, placebo-controlled trial of ruxolitinib for myelofibrosis. *N Engl J Med, 366*(9), 799-807. doi:10.1056/NEJMoa1110557

Walton, M. K., Powers, J. H., 3rd, Hobart, J., Patrick, D., Marquis, P., Vamvakas, S., . . . Burke, L. B. (2015). Clinical Outcome Assessments: Conceptual Foundation-Report of the ISPOR Clinical Outcomes Assessment - Emerging Good Practices for Outcomes Research Task Force. *Value Health, 18*(6), 741-752. doi:10.1016/j.jval.2015.08.006

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research, 11*, 3571-3594.

Weenink, J. W., Braspenning, J., & Wensing, M. (2014). Patient reported outcome measures (PROMs) in primary care: an observational pilot study of seven generic instruments. *BMC family practice, 15*, 88.

Zagadailov, E., Fine, M., & Shields, A. (2013). Patient-reported outcomes are changing the landscape in oncology care: challenges and opportunities for payers. *Am Health Drug Benefits*, *6*(5), 264-274. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24991362

Zoomerang.com. Web-based survey tool. Retrieved from http://www.zoomerang.com/ .

Appendix

Additional Simulation and/or Application Results for Chapters Two and Three

Table S2.1. Percent of CFA simulation iterations that fail to converge and/or produce out of bound item-to-domain correlation (i.e., $\boldsymbol{\rho_j \notin [-1, 1]}$).

Number of Items (P)	Number of Participants (N)	Number of Response Categories (C)	CFA Fail to Converge (%)	CFA Out of Bound Estimate (%)
4	50	2	6	21
	50	5	1	13
	50	7	0	14
	100	2	3	14
	100	5	0	3
	100	7	1	4
	200	2	2	5
	200	5	0	1
	200	7	0	1
	500	2	0	1
	500	5	0	0
	500	7	0	0
6	50	2	2	21
	50	5	0	2
	50	7	0	2
	100	2	0	3
	100	5	0	0
	100	7	0	1
	200	2	0	2
	200	5	0	0
	200	7	0	0
	500	2	0	0
	500	5	0	0
	500	7	0	0
9	50	2	0	6
	50	5	0	0
	50	7	0	0
	100	2	0	0
	100	5	0	0
	100	7	0	0
	200	2	0	0
	200	5	0	0
	200	7	0	0
	500	2	0	0
	500	5	0	0
	500	7	0	0

Table S2.2. Item-to-domain correlation $\boldsymbol{\rho}$ estimates and standard errors for prior (content experts), OBID posterior informative (experts information used), and OBID posterior non-informative (experts information not used).

		Hispanic (N=36)		African American (N=34)	
Item	Expert Prior	OBID (Posterior Informative)	OBID (Posterior Non-informative)	OBID (Posterior Informative)	OBID (Posterior Non-informative)
Item 1	0.381 (0.130)	0.466 (0.093)	0.710 (0.123)	0.495 (0.086)	0.774 (0.102)
Item 2	0.673 (0.112)	0.565 (0.118)	0.570 (0.160)	0.674 (0.088)	0.791 (0.094)
Item 3	0.472 (0.119)	0.615 (0.074)	0.914 (0.055)	0.653 (0.066)	0.942 (0.036)
Item 4	0.629 (0.109)	0.717 (0.070)	0.920 (0.053)	0.718 (0.068)	0.884 (0.059)
Item 5	0.528 (0.116)	0.537 (0.097)	0.607 (0.159)	0.641 (0.074)	0.908 (0.056)
Item 6	0.562 (0.110)	0.647 (0.079)	0.783 (0.110)	0.620 (0.077)	0.819 (0.079)
Item 7	0.561 (0.118)	0.653 (0.082)	0.784 (0.110)	0.725 (0.062)	0.938 (0.037)

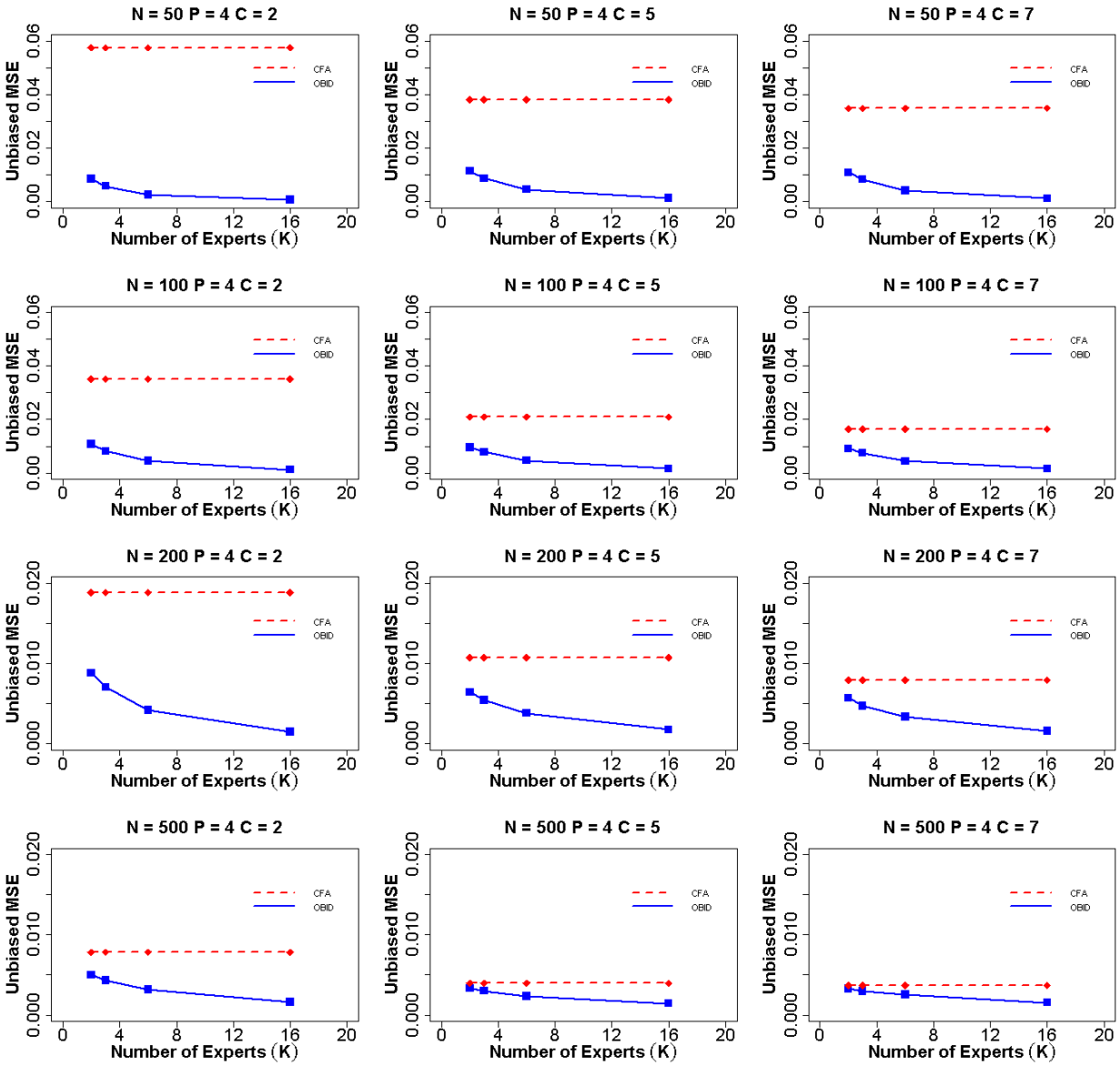


Figure S2.1. Average MSE of item-to-domain correlation ρ for four items and unbiased experts. Average mean squared error (MSE) for item-to-domain correlation ρ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 4$ (number of items) and experts are unbiased $\{\rho_0 = (0.50, 0.30, 0.70, 0.50)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

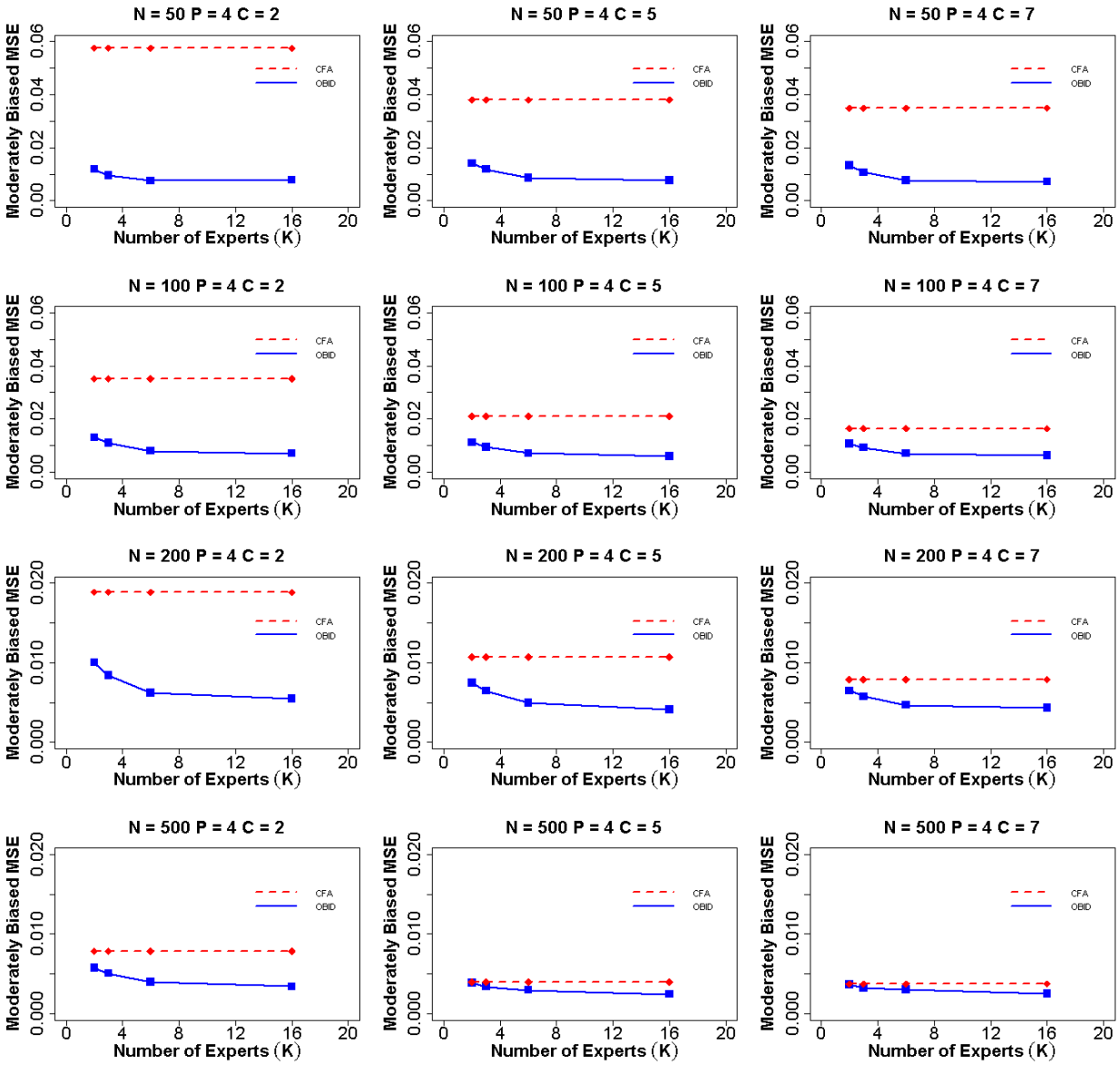


Figure S2.2. Average MSE of item-to-domain correlation ρ for four items and moderately biased experts. Average mean squared error (MSE) for item-to-domain correlation ρ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 4$ (number of items) and experts are moderately biased $\{\rho_0 = (0.60, 0.40, 0.80, 0.60)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.
Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

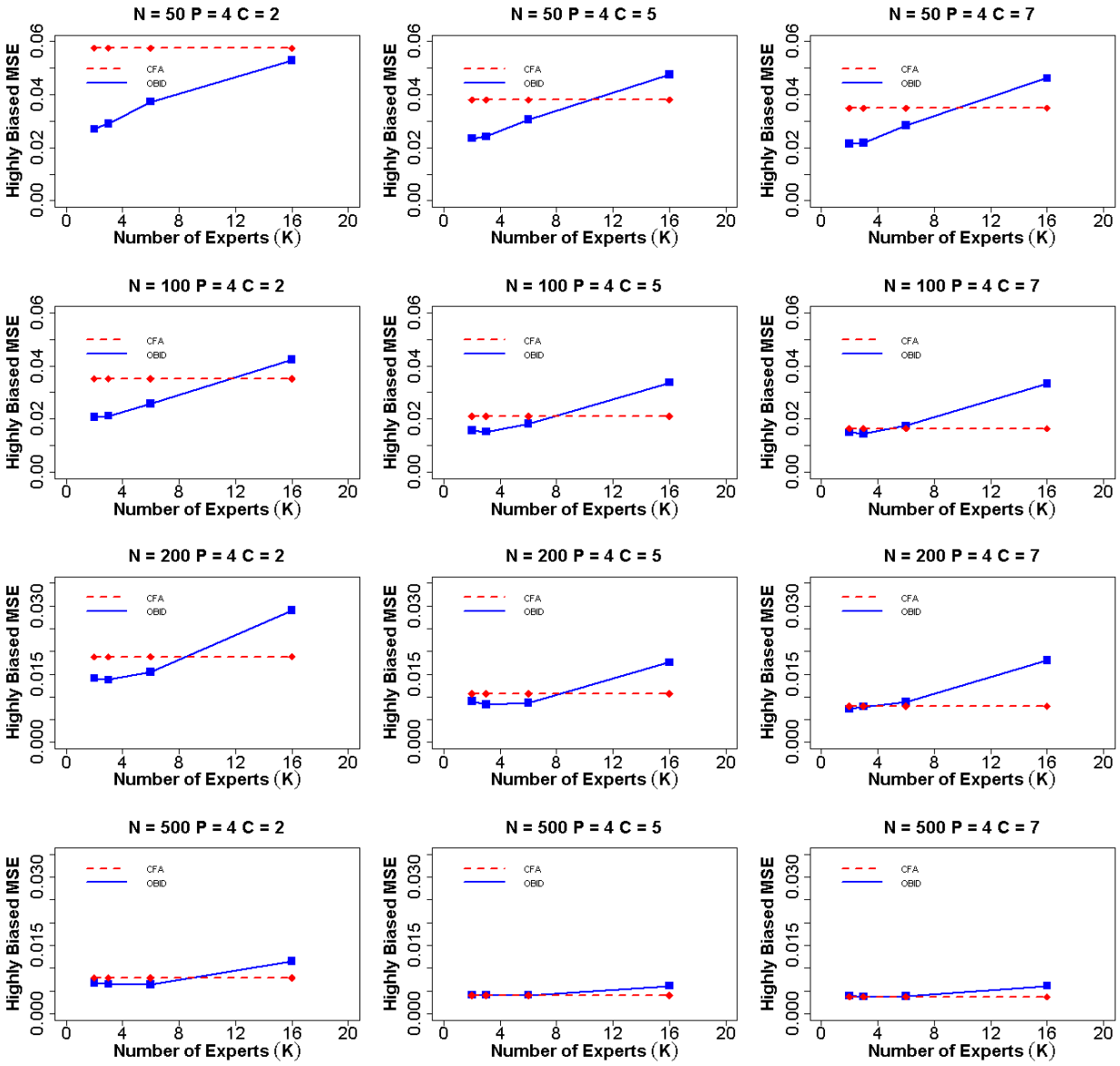


Figure S2.3. Average MSE of item-to-domain correlation ρ for four items and highly biased experts. Average mean squared error (MSE) for item-to-domain correlation ρ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 4$ (number of items) and experts are highly biased $\{\rho_0 = (0.75, 0.65, 0.85, 0.75)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

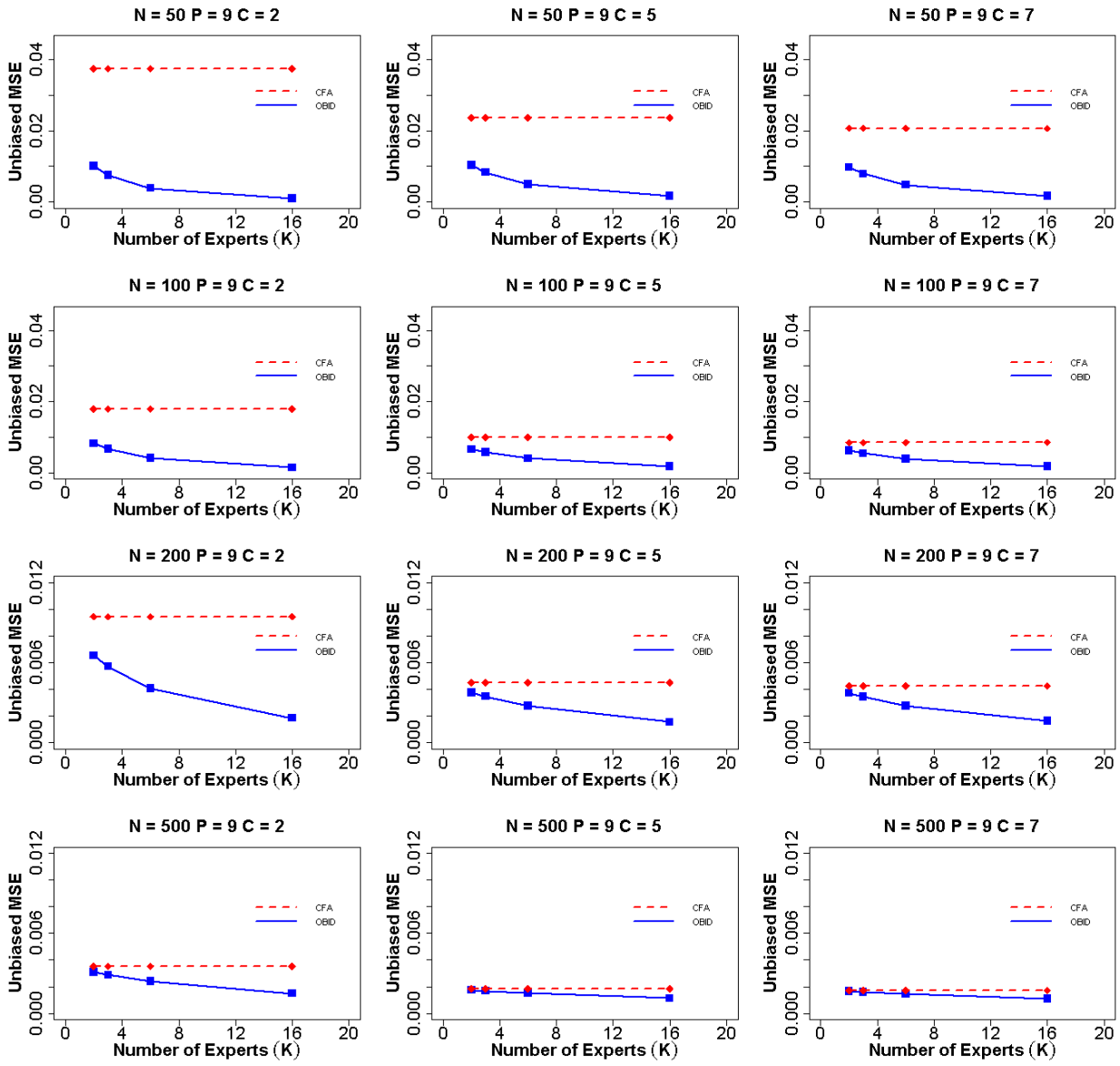


Figure S2.4. Average MSE of item-to-domain correlation ρ for nine items and unbiased experts. Average mean squared error (MSE) for item-to-domain correlation ρ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 9$ (number of items) and experts are unbiased $\{\rho_0 = (0.30, 0.50, 0.70, 0.70, 0.30, 0.50, 0.70, 0.50, 0.30)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

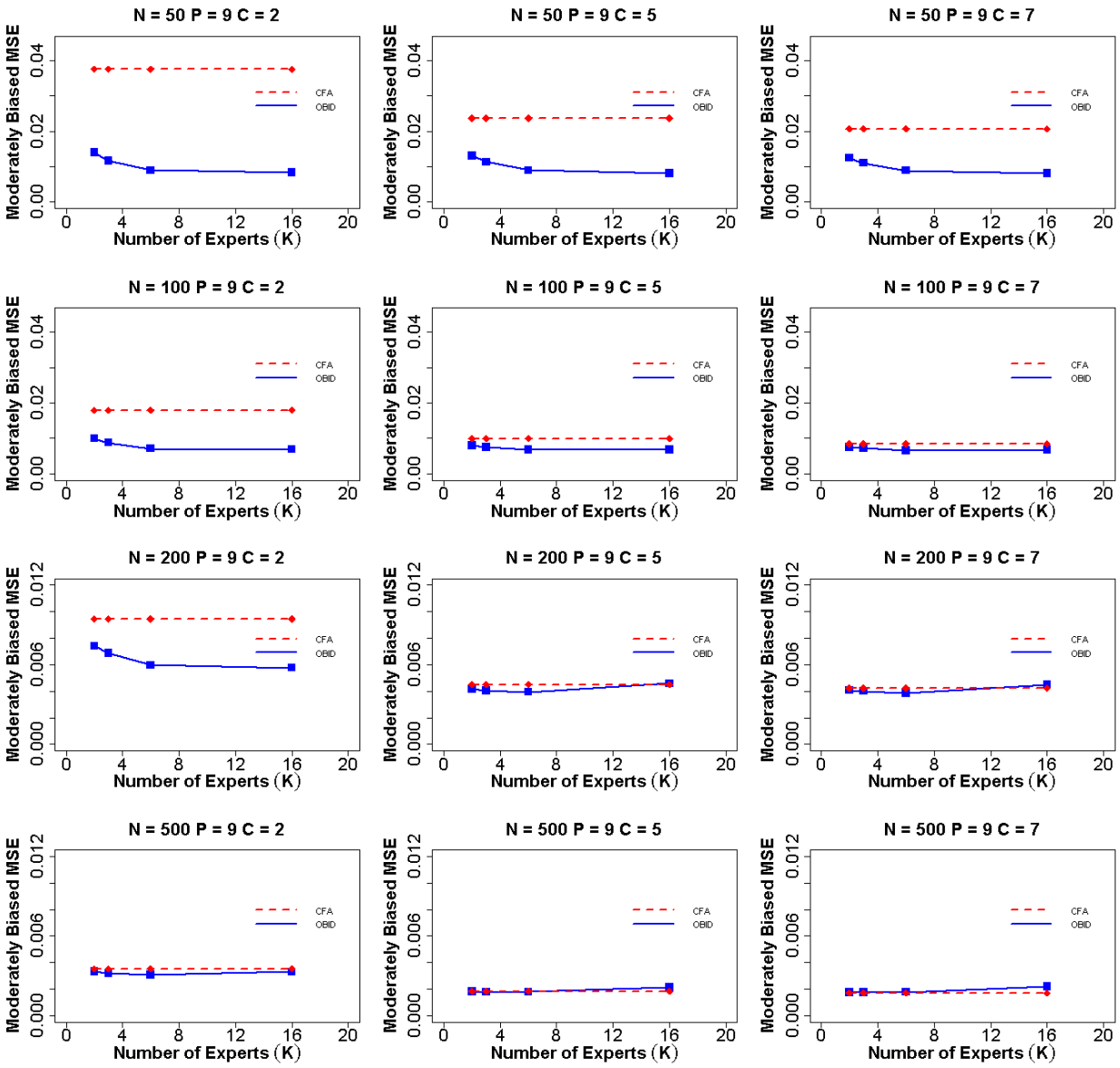


Figure S2.5. Average MSE of item-to-domain correlation ρ for nine items and moderately biased experts. Average mean squared error (MSE) for item-to-domain correlation ρ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 9$ (number of items) and experts are moderately biased $\{\rho_0 = (0.40, 0.60, 0.80, 0.80, 0.40, 0.60, 0.80, 0.60, 0.40)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

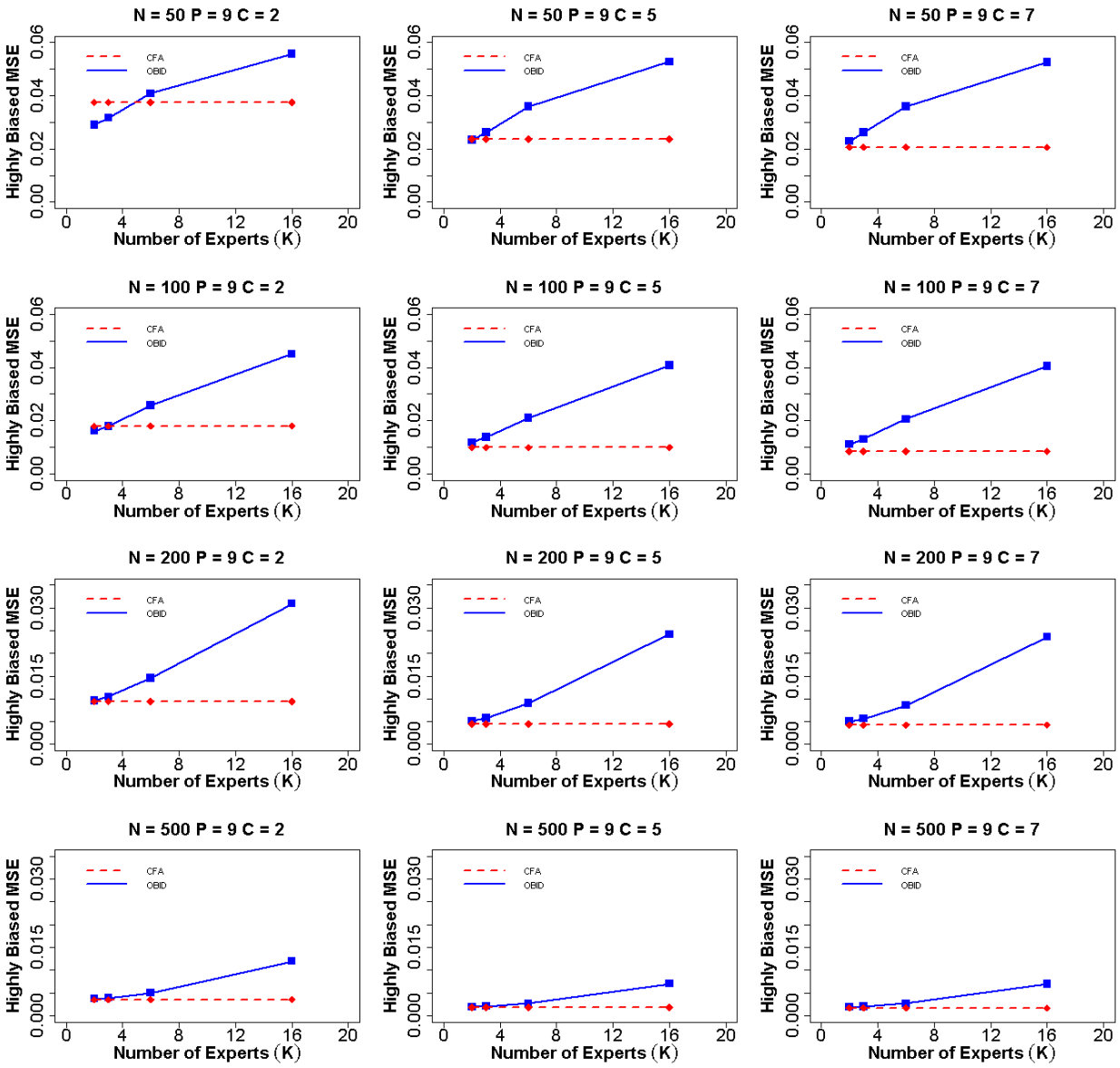


Figure S2.6. Average MSE of item-to-domain correlation ρ for nine items and highly biased experts. Average mean squared error (MSE) for item-to-domain correlation ρ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 9$ (number of items) and experts are highly biased $\{\rho_0 = (0.65,0.75,0.85,0.85,0.65,0.75,0.85,0.75,0.65)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

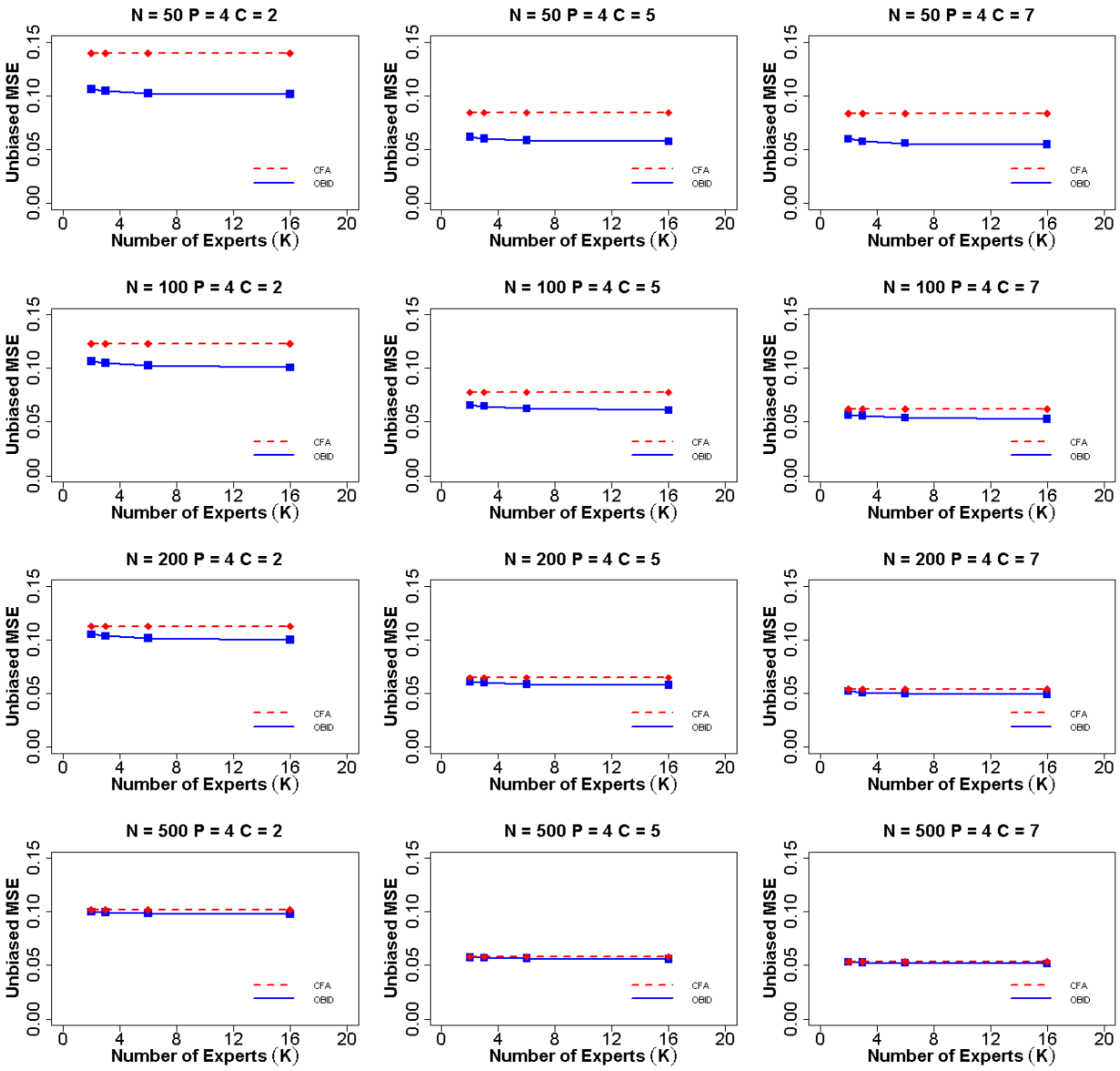


Figure S2.7. Average MSE of validity coefficient γ for four items and unbiased experts. Mean squared error (MSE) for validity coefficient γ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 4$ (number of items) and experts are unbiased $\{\rho_0 = (0.50, 0.30, 0.70, 0.50)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

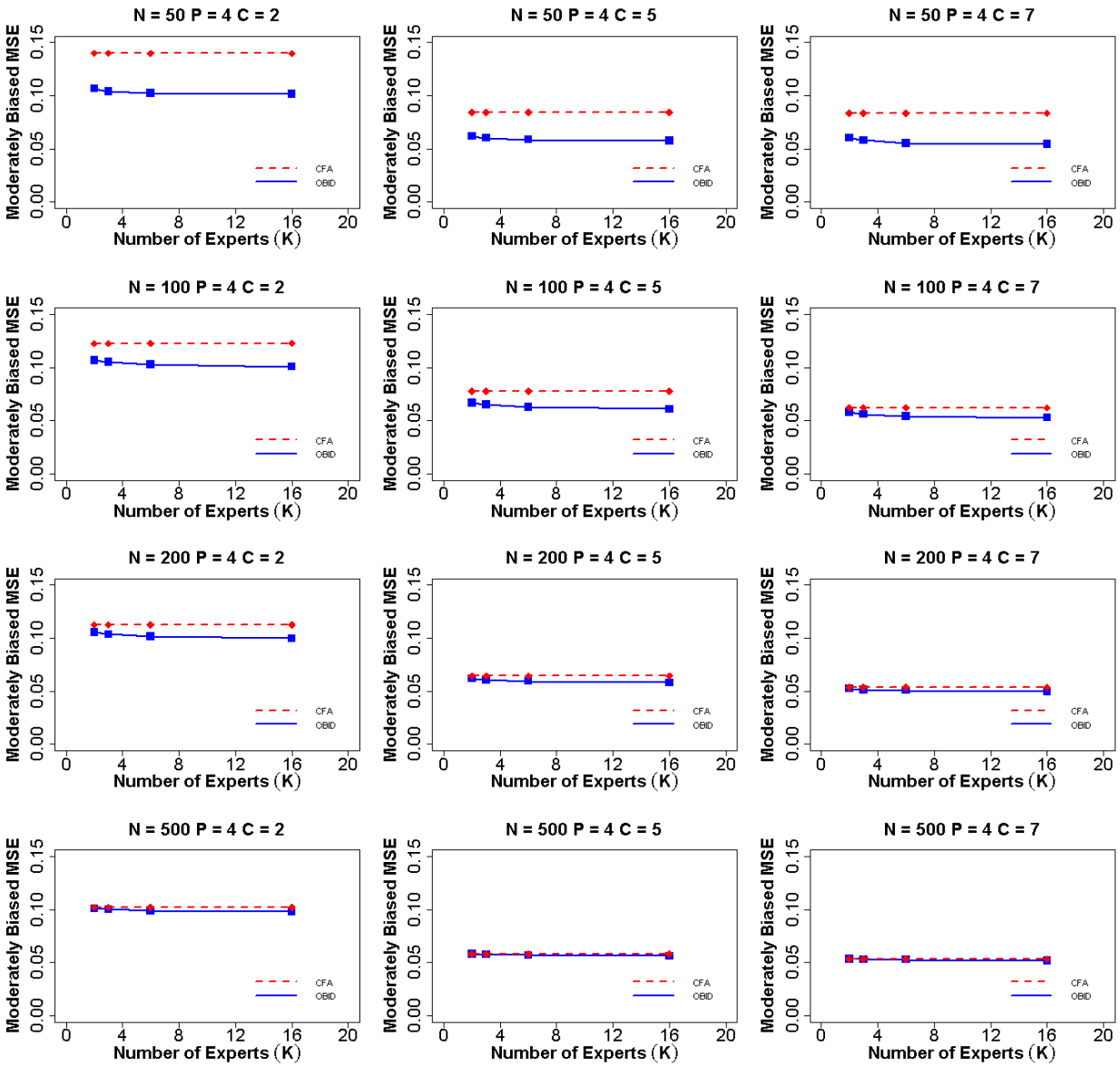


Figure S2.8. Average MSE of validity coefficient γ for four items and moderately biased experts. Average mean squared error (MSE) for validity coefficient γ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 4$ (number of items) and experts are moderately biased $\{\rho_0 = (0.60, 0.40, 0.80, 0.60)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

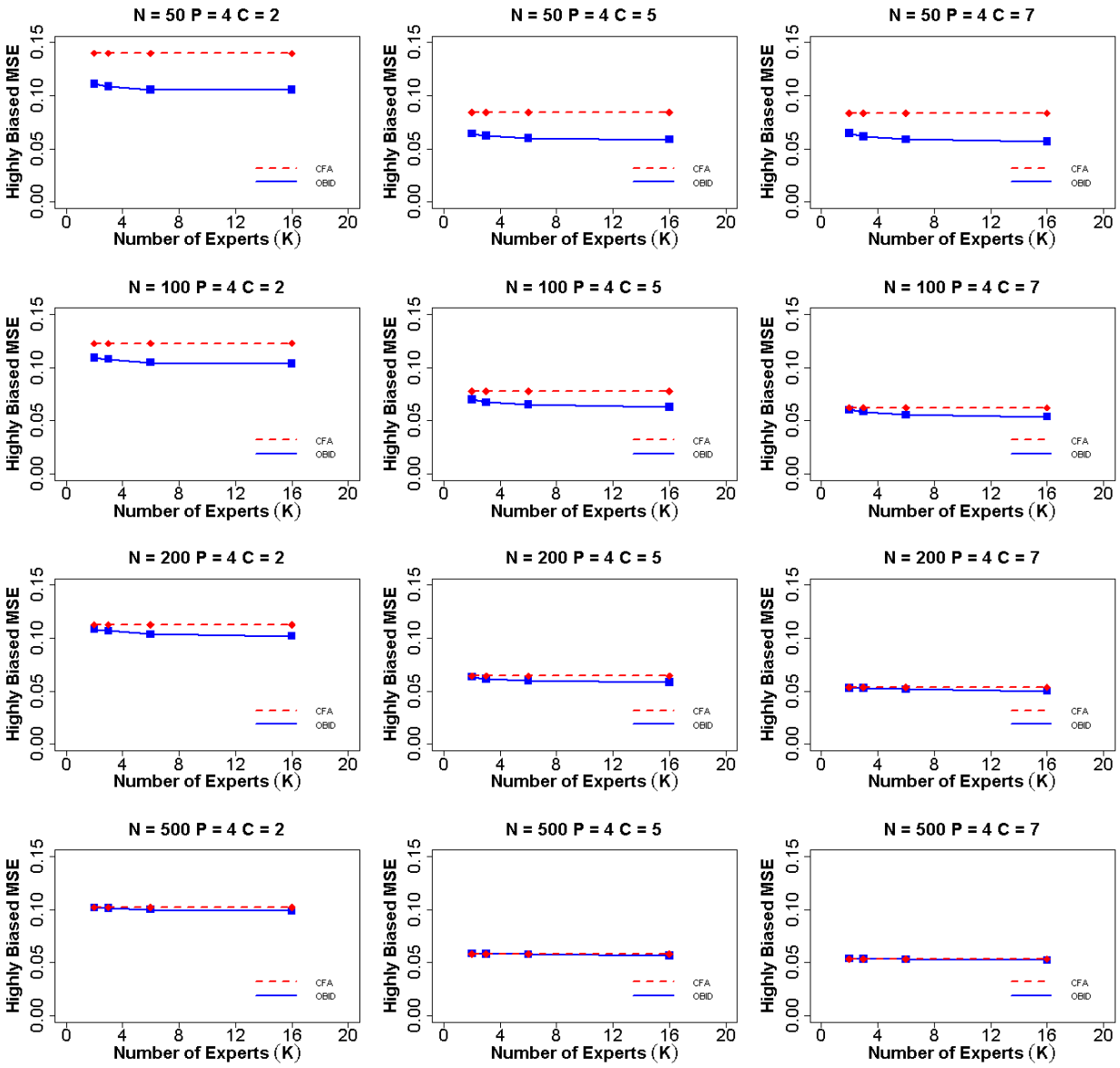


Figure S2.9. Average MSE of validity coefficient γ for four items and highly biased experts. Average mean squared error (MSE) for validity coefficient γ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 4$ (number of items) and experts are highly biased $\{\rho_0 = (0.75, 0.65, 0.85, 0.75)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

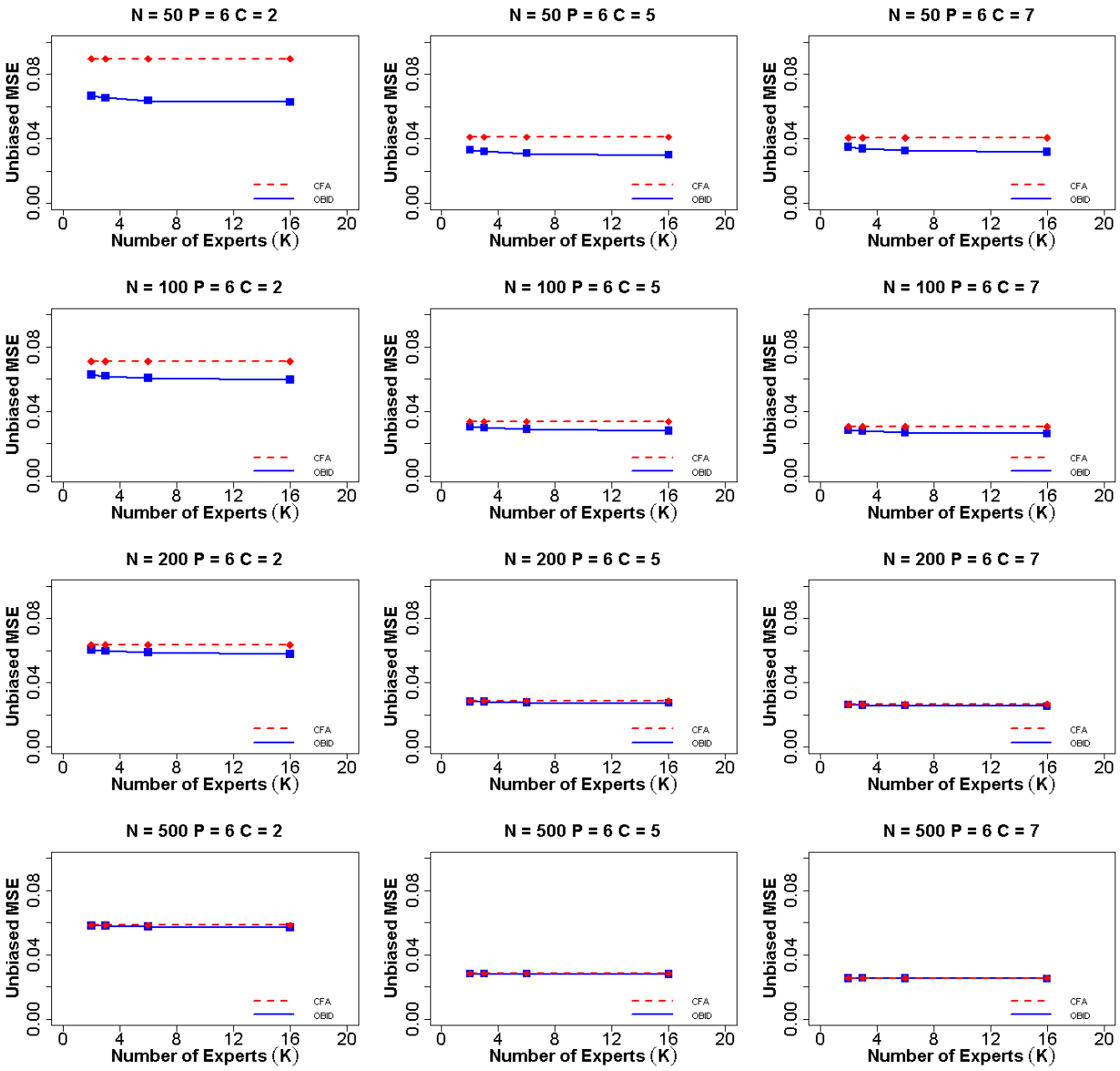


Figure S2.10. Average MSE of validity coefficient γ for six items and unbiased experts. Average mean squared error (MSE) for validity coefficient γ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 6$ (number of items) and experts are unbiased $\{\rho_0 = (0.30, 0.50, 0.70, 0.70, 0.30, 0.50)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

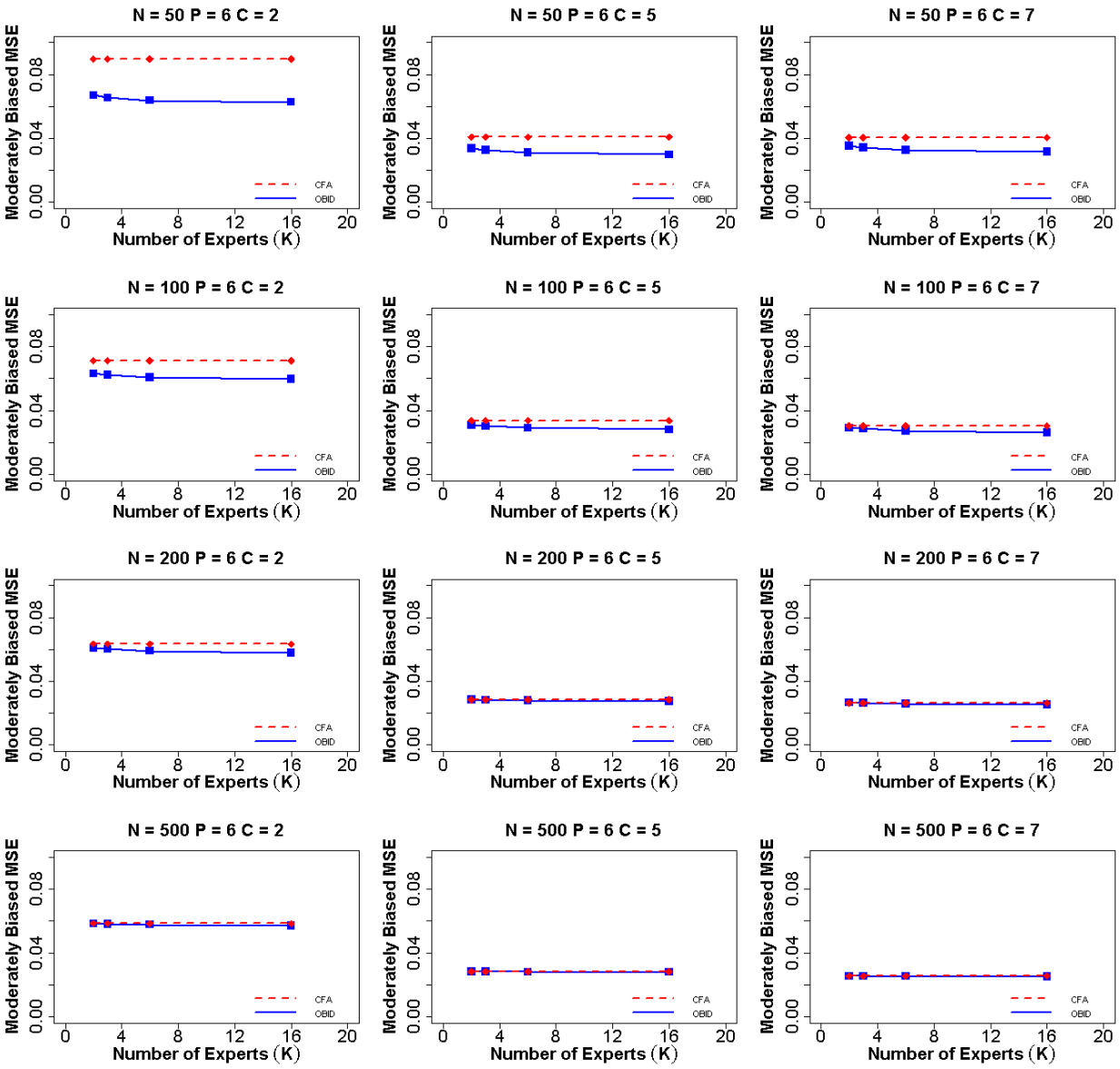


Figure S2.11. Average MSE of validity coefficient γ for six items and moderately biased experts. Average mean squared error (MSE) for validity coefficient γ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 6$ (number of items) and experts are moderately biased $\{\rho_0 = (0.40, 0.60, 0.80, 0.80, 0.40, 0.60)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

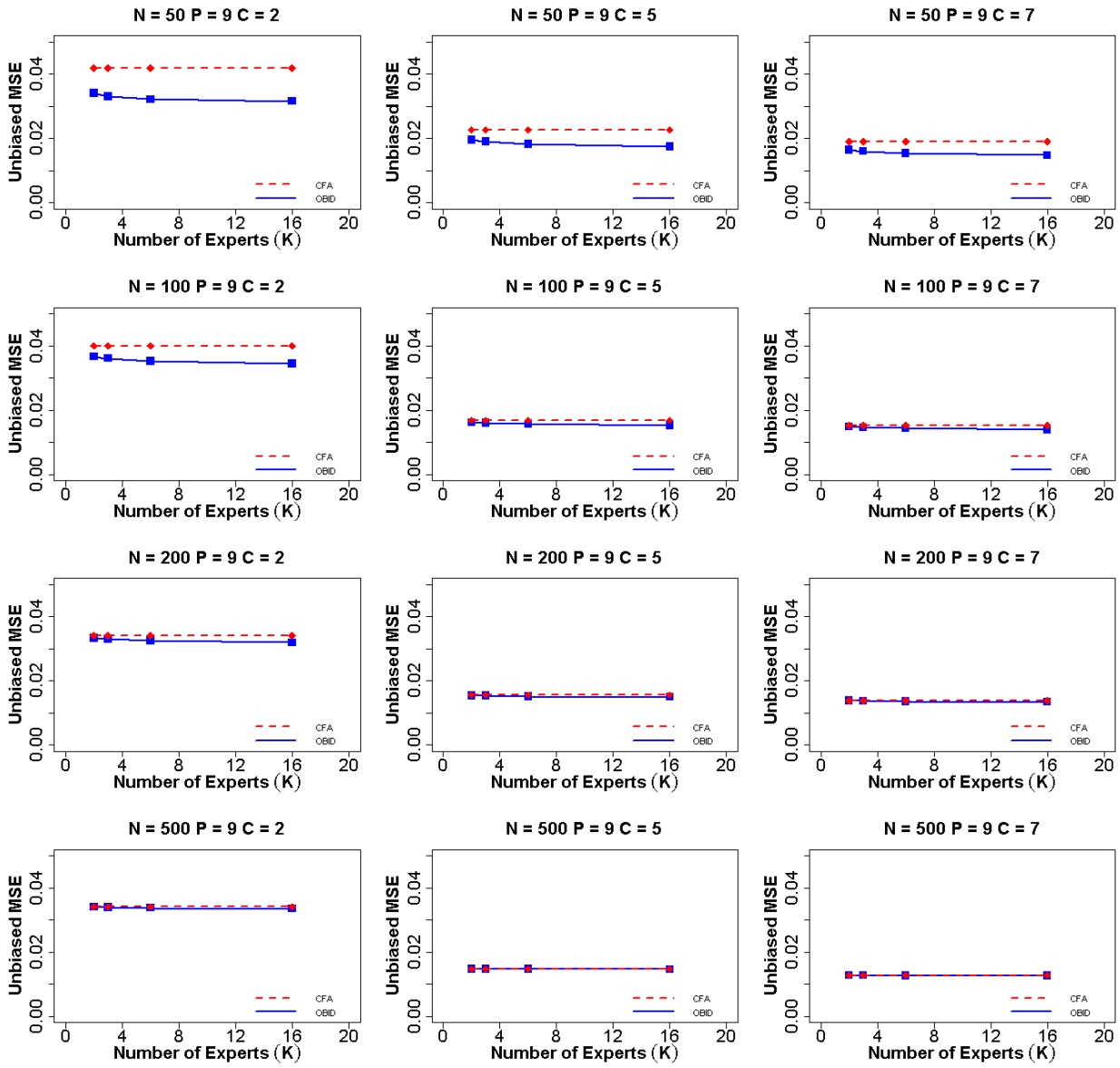


Figure S2.12. Average MSE of validity coefficient γ for nine items and unbiased experts. Average mean squared error (MSE) for validity coefficient γ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 9$ (number of items) and experts are unbiased $\{\rho_0 = (0.30, 0.50, 0.70, 0.70, 0.30, 0.50, 0.70, 0.50, 0.30)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

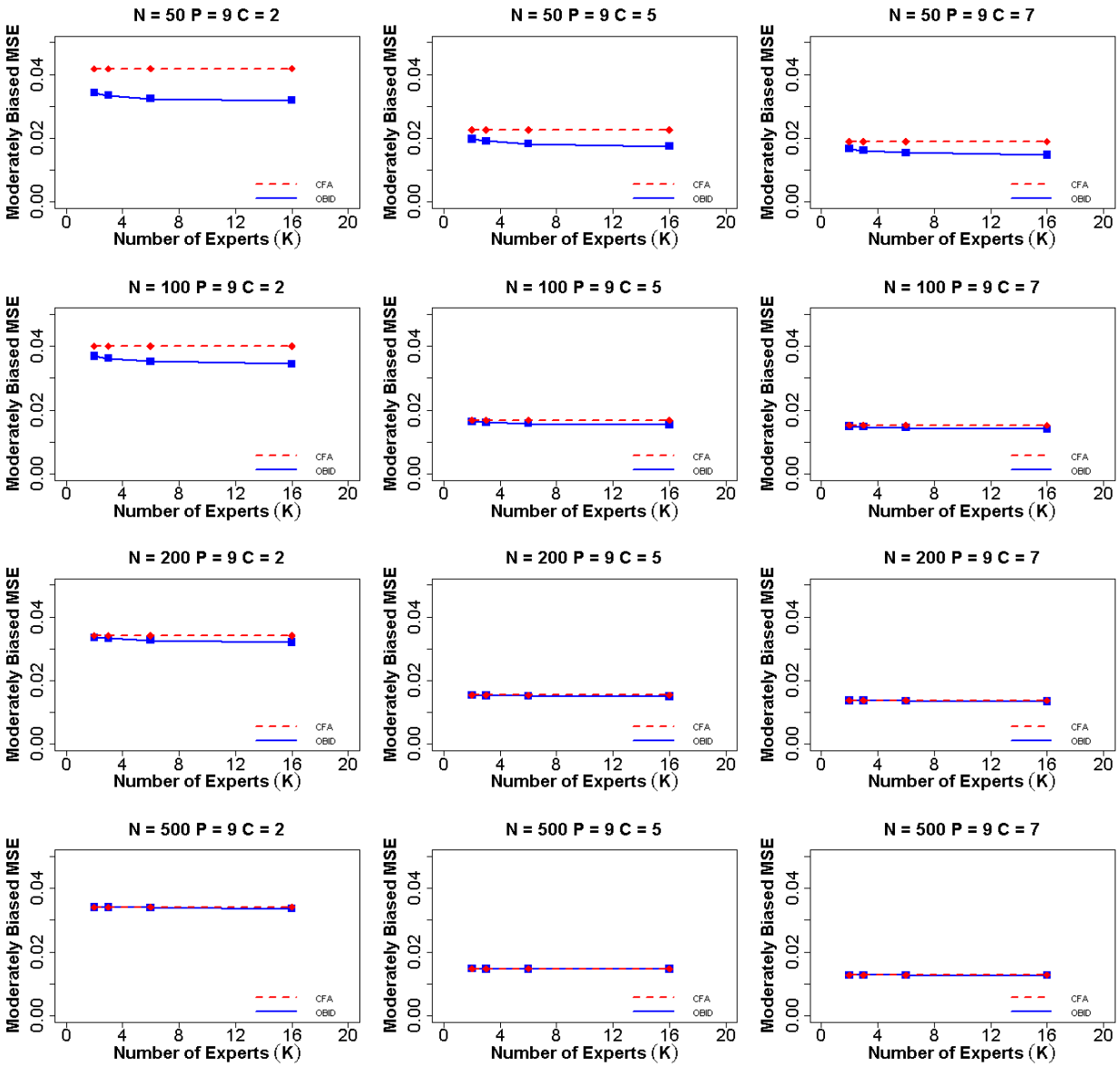


Figure S2.13. Average MSE of validity coefficient γ for nine items and moderately biased experts. Average mean squared error (MSE) for validity coefficient γ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 9$ (number of items) and experts are moderately biased $\{\rho_0 = (0.40, 0.60, 0.80, 0.80, 0.40, 0.60, 0.80, 0.60, 0.40)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

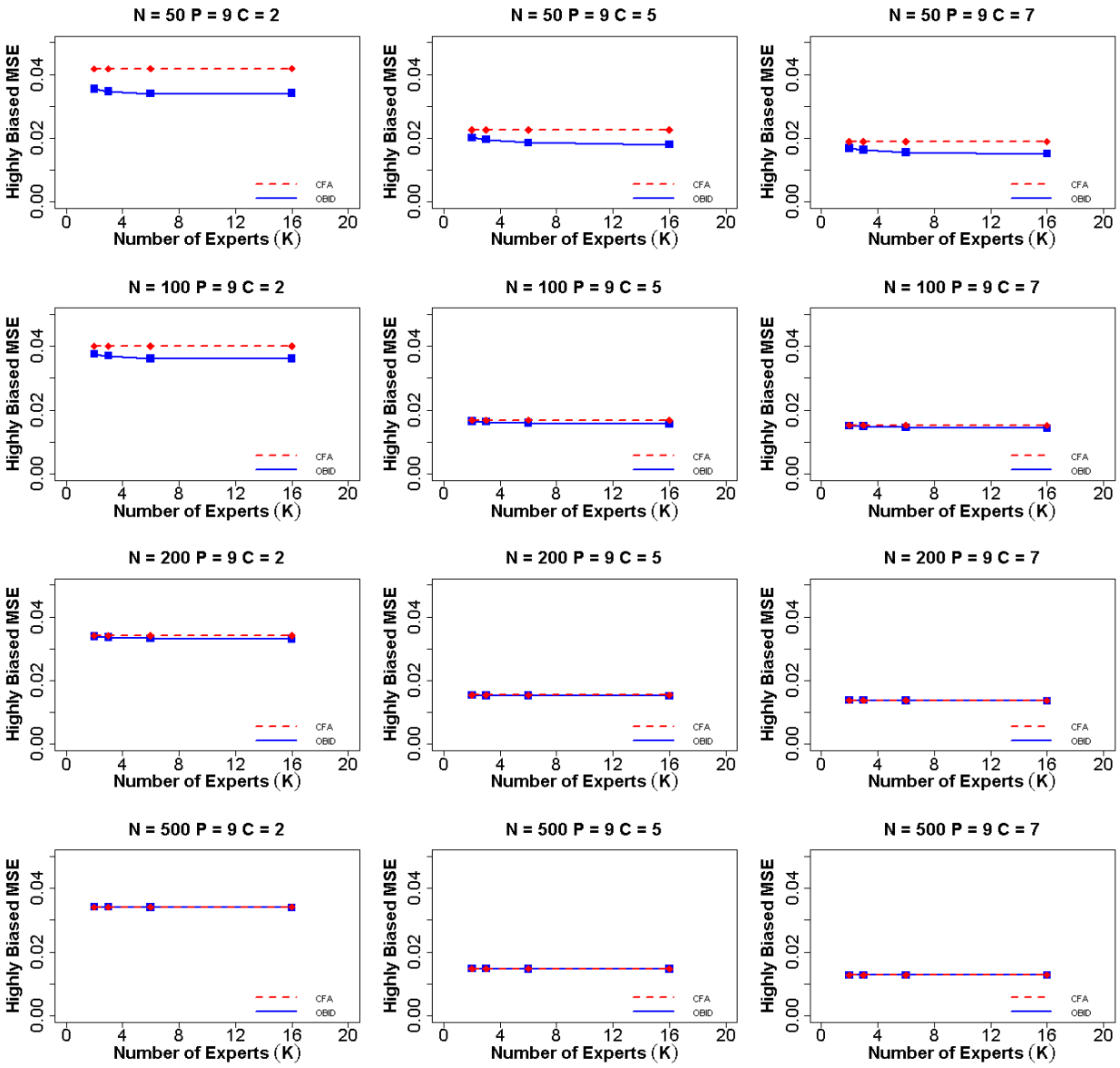


Figure S2.14. Average MSE of validity coefficient γ for nine items and highly biased experts. Average mean squared error (MSE) for validity coefficient γ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 9$ (number of items) and experts are highly biased $\{\rho_0 = (0.65,0.75,0.85,0.85,0.65,0.75,0.85,0.75,0.65)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

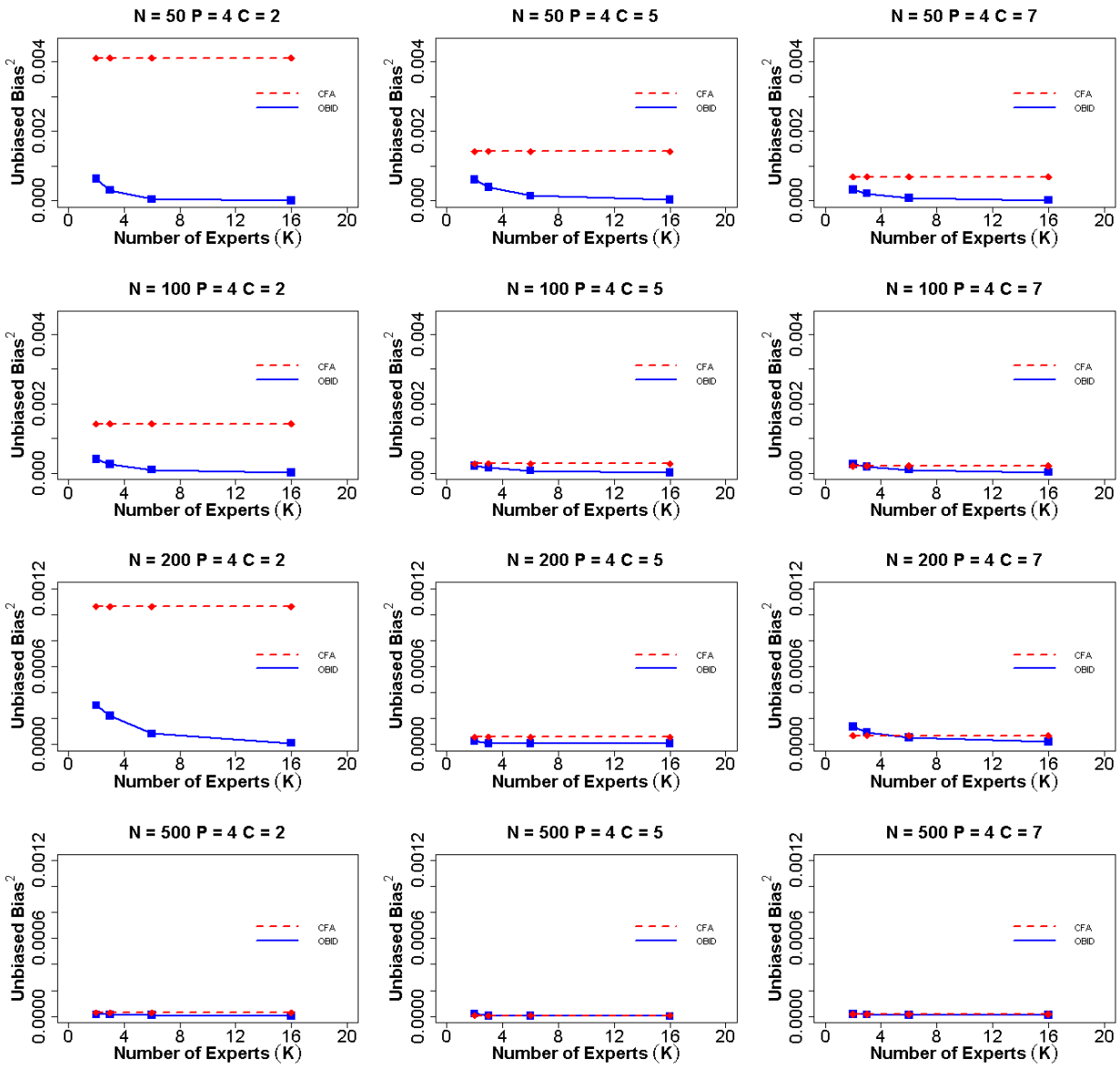


Figure S2.15. Average squared bias for item-to-domain correlation ρ for four items and unbiased experts. Average squared bias for item-to-domain correlation ρ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 4$ (number of items) and experts are unbiased $\{\rho_0 = (0.50, 0.30, 0.70, 0.50)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

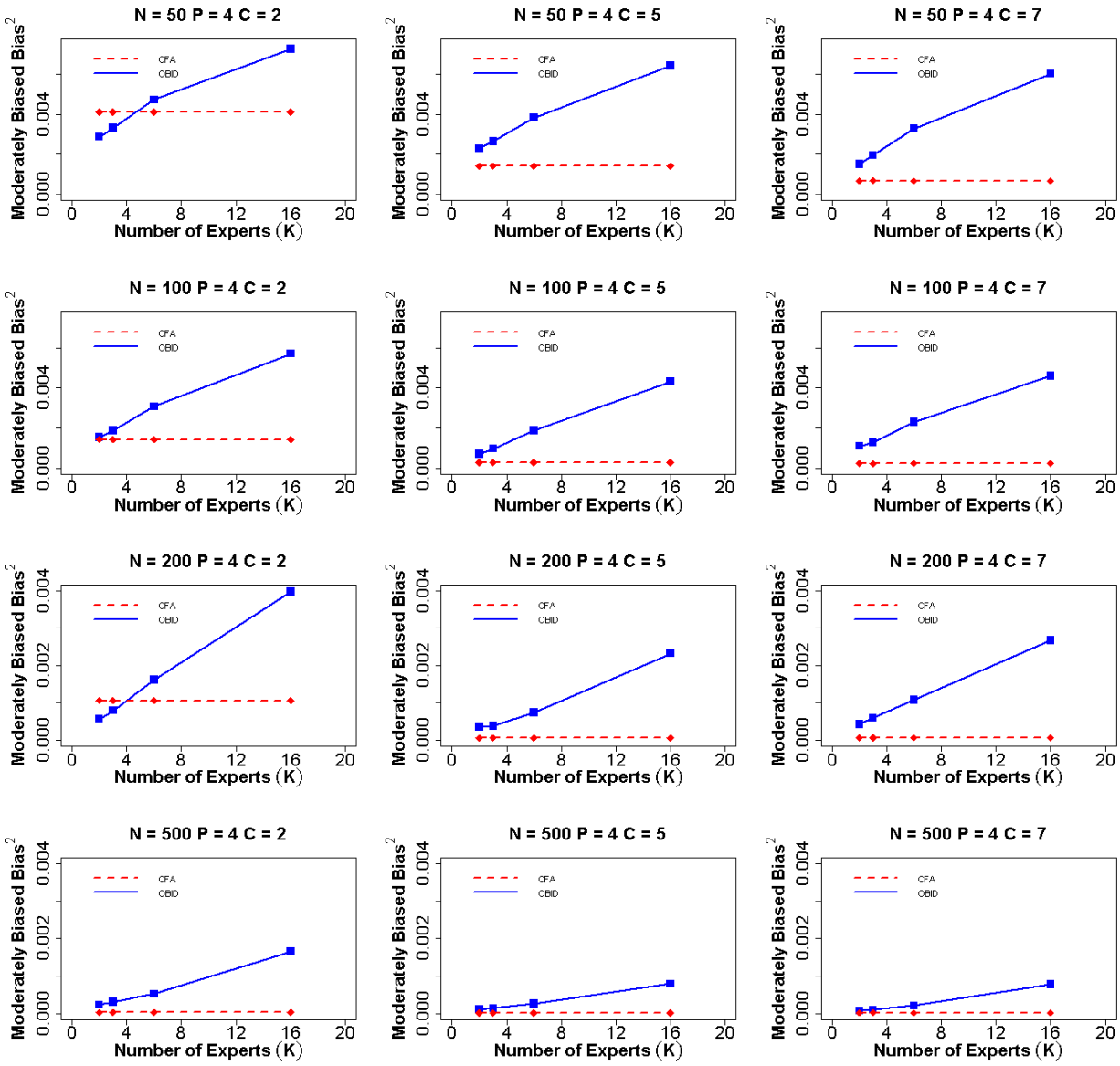


Figure S2.16. Average squared bias for item-to-domain correlation ρ for four items and moderately biased experts. Average squared bias for item-to-domain correlation ρ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 4$ (number of items) and experts are moderately biased $\{\rho_0 = (0.60, 0.40, 0.80, 0.60)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

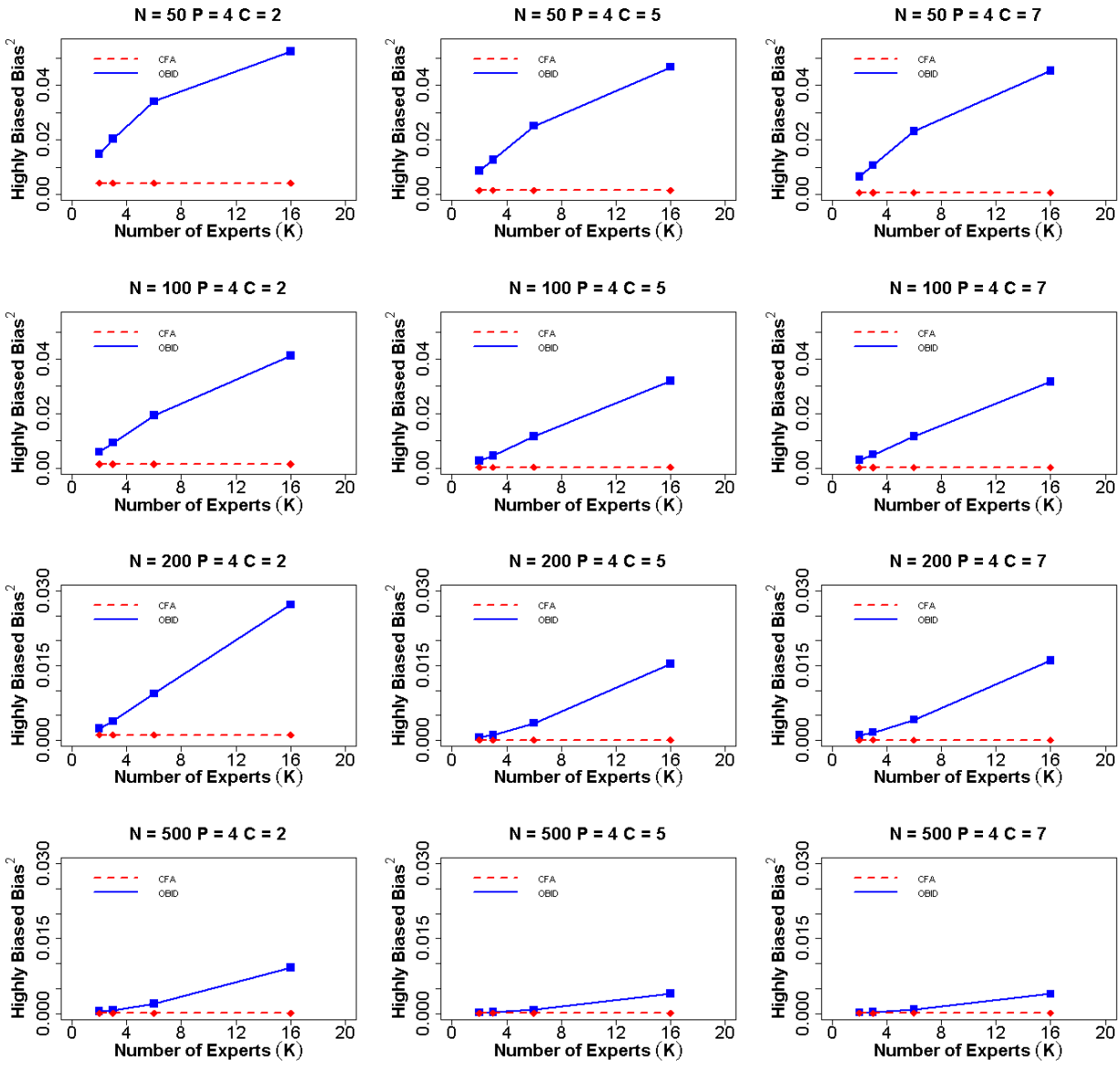


Figure S2.17. Average squared bias for item-to-domain correlation ρ for four items and highly biased experts. Average squared bias for item-to-domain correlation ρ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 4$ (number of items) and experts are highly biased $\{\rho_0 = (0.75, 0.65, 0.85, 0.75)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

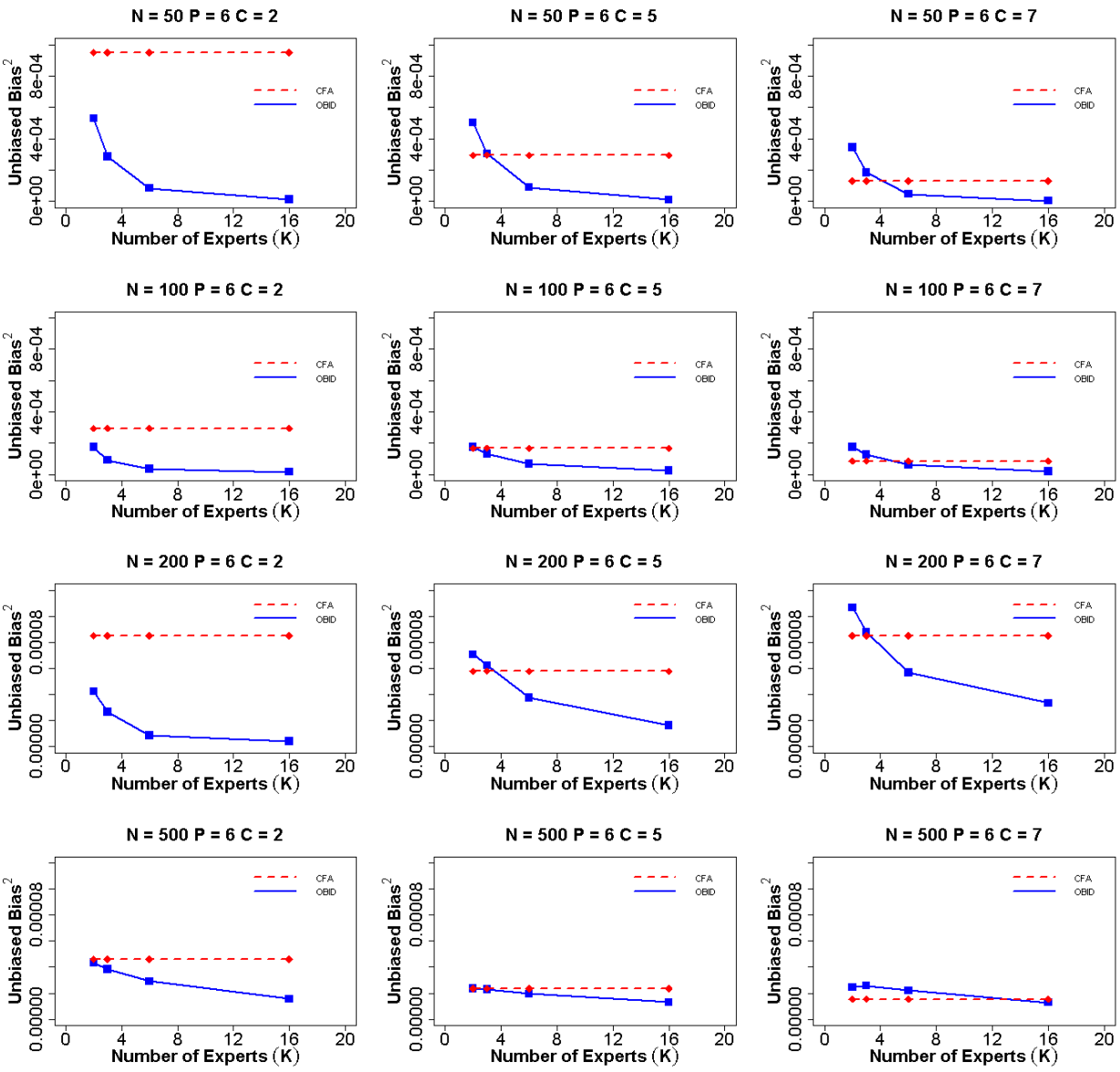


Figure S2.18. Average squared bias for item-to-domain correlation ρ for six items and unbiased experts. Average squared bias for item-to-domain correlation ρ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 6$ (number of items) and experts are unbiased $\{\rho_0 = (0.30, 0.50, 0.70, 0.70, 0.30, 0.50)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

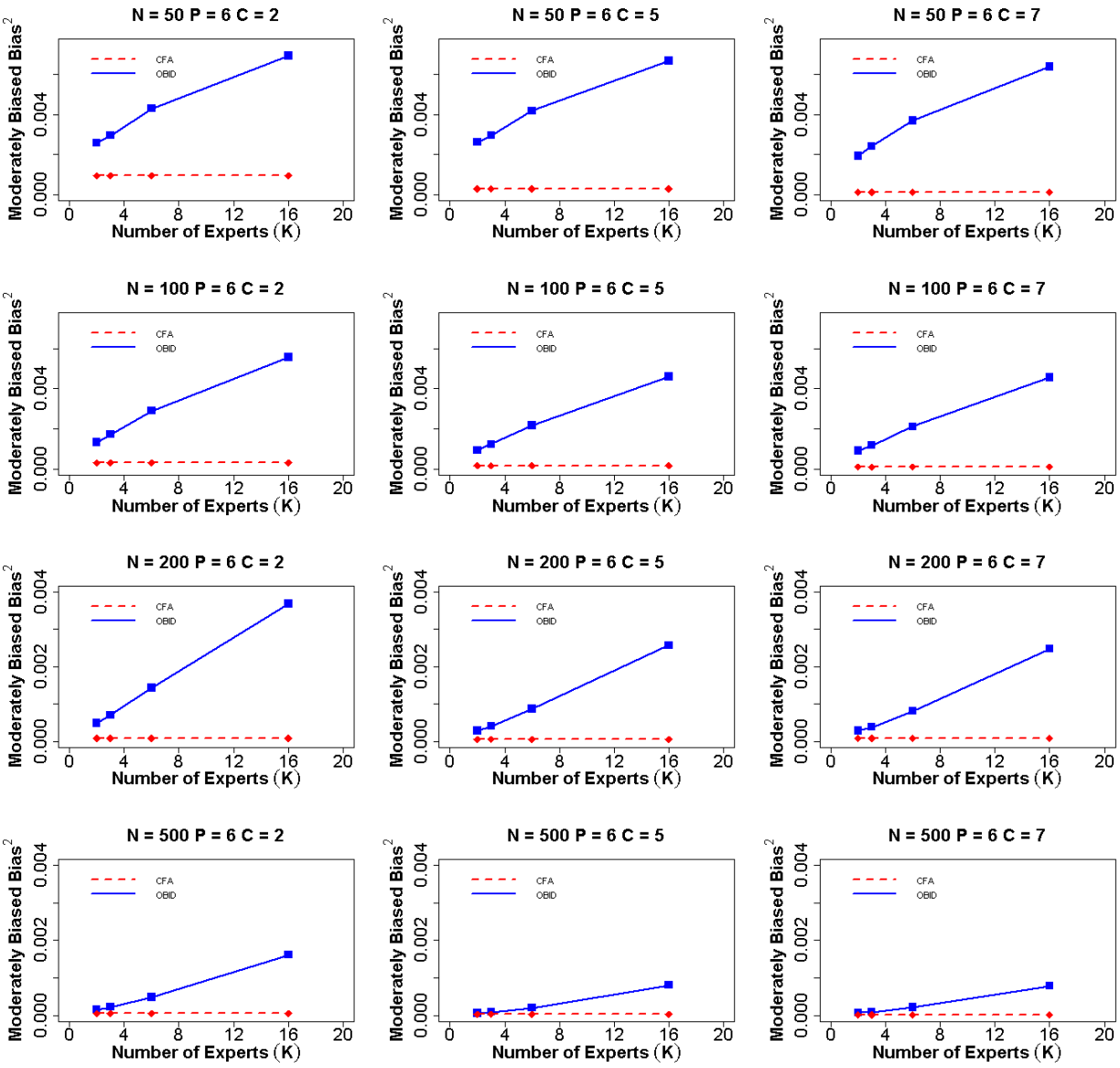


Figure S2.19. Average squared bias for item-to-domain correlation ρ for six items and moderately biased experts. Average squared bias for item-to-domain correlation ρ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 6$ (number of items) and experts are moderately biased $\{\rho_0 = (0.40, 0.60, 0.80, 0.80, 0.40, 0.60)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

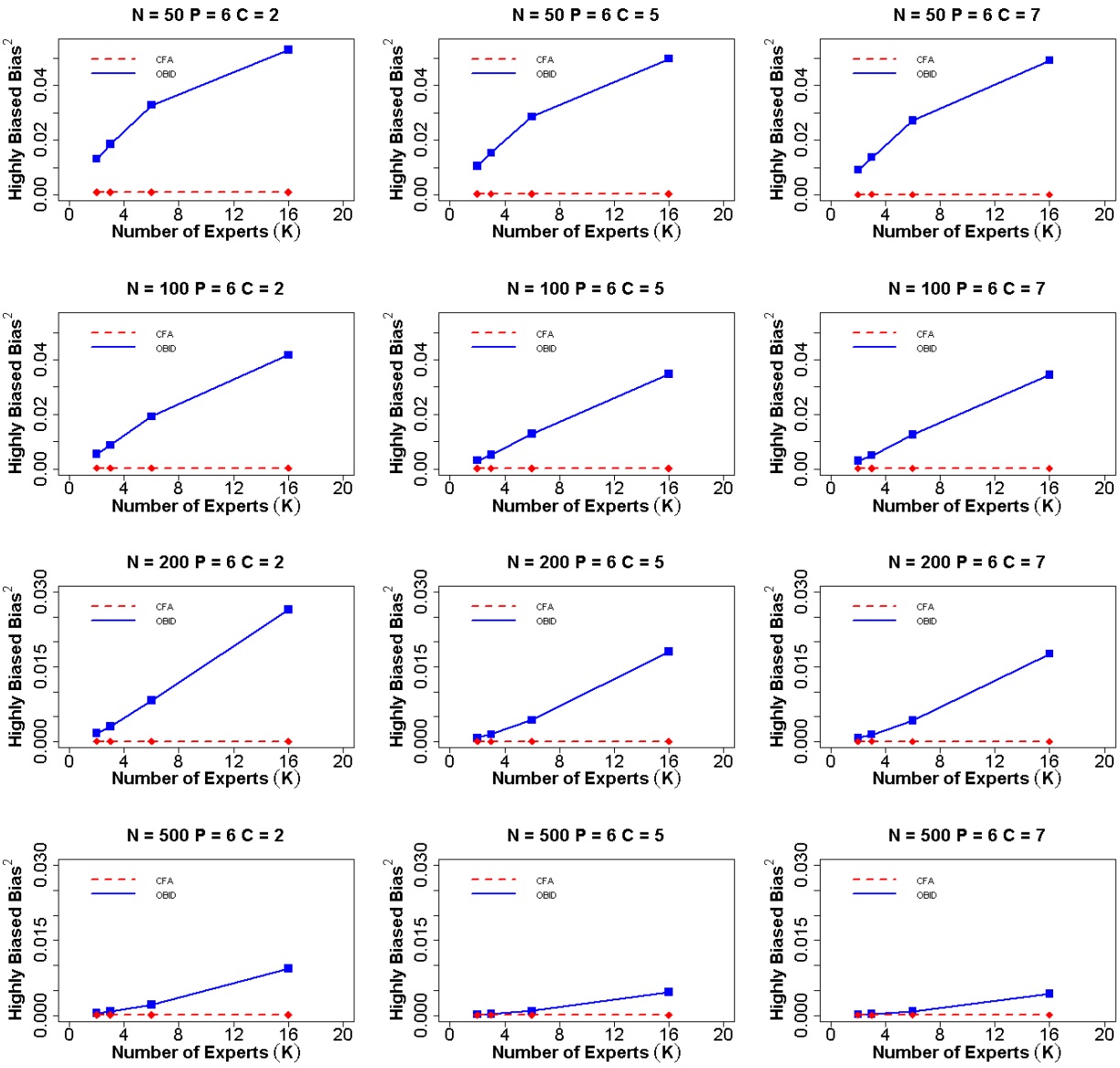


Figure S2.20. Average squared bias for item-to-domain correlation ρ for six items and highly biased experts. Average squared bias for item-to-domain correlation ρ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 6$ (number of items) and experts are highly biased $\{\rho_0 = (0.65, 0.75, 0.85, 0.85, 0.65, 0.75)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

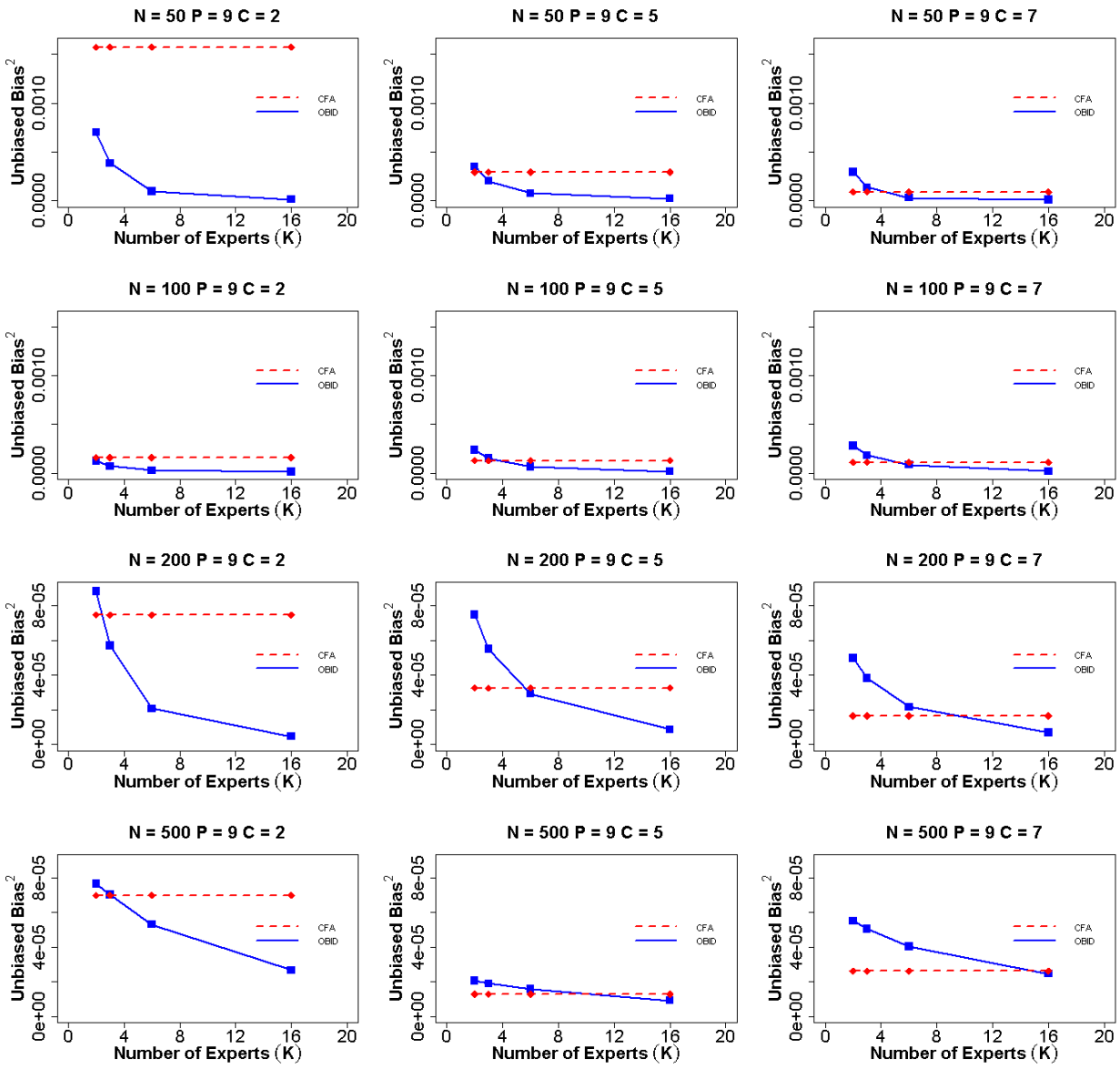


Figure S2.21. Average squared bias for item-to-domain correlation ρ for nine items and unbiased experts. Average squared bias for item-to-domain correlation ρ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 9$ (number of items) and experts are unbiased $\{\rho_0 = (0.30, 0.50, 0.70, 0.70, 0.30, 0.50, 0.70, 0.50, 0.30)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

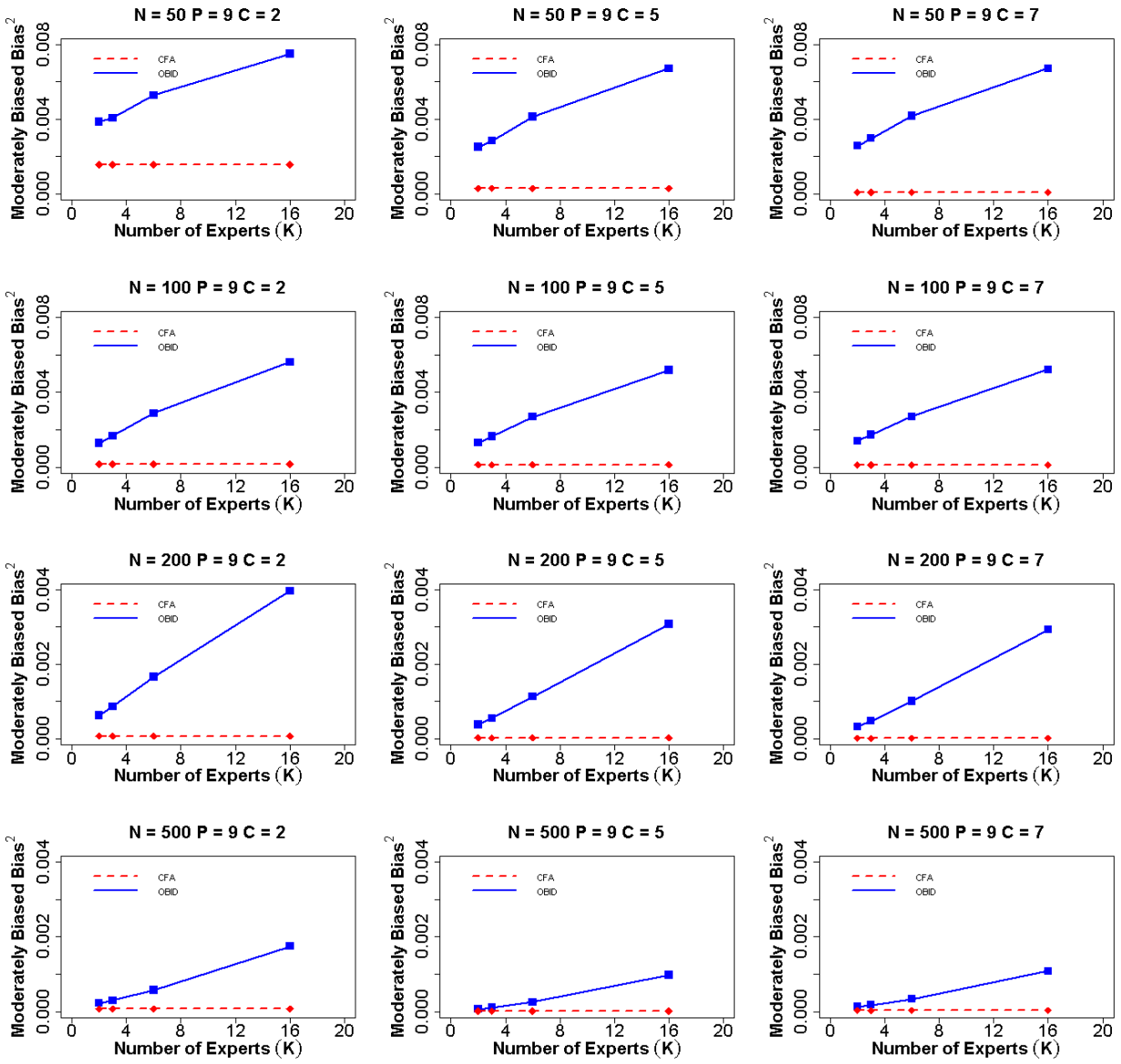


Figure S2.22. Average squared bias for item-to-domain correlation ρ for nine items and moderately biased experts. Average squared bias for item-to-domain correlation ρ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 9$ (number of items) and experts are moderately biased $\{\rho_0 = (0.40, 0.60, 0.80, 0.80, 0.40, 0.60, 0.80, 0.60, 0.40)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

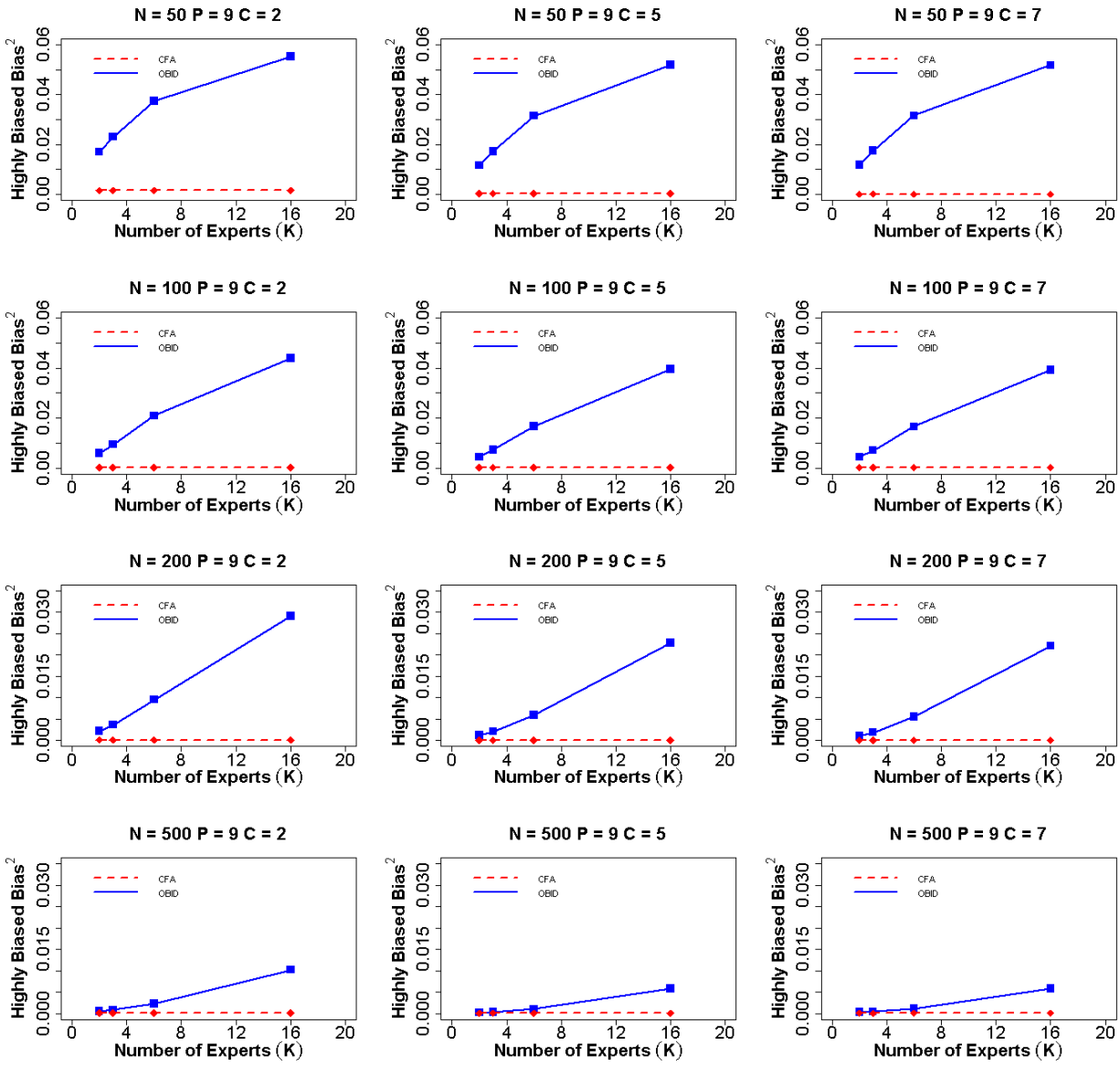


Figure S2.23. Average squared bias for item-to-domain correlation ρ for nine items and highly biased experts. Average squared bias for item-to-domain correlation ρ using OBID (solid blue line) and ordinal CFA (dashed red line) when $P = 9$ (number of items) and experts are highly biased $\{\rho_0 = (0.65,0.75,0.85,0.85,0.65,0.75,0.85,0.75,0.65)\}$. The participant sample sizes are N = 50, 100, 200, and 500. The numbers of response categories are C = 2, 5, and 7, and the numbers of experts are K = 2, 3, 6, and 16.

Note. OBID = Ordinal Bayesian Instrument Development; CFA = Confirmatory Factor Analysis.

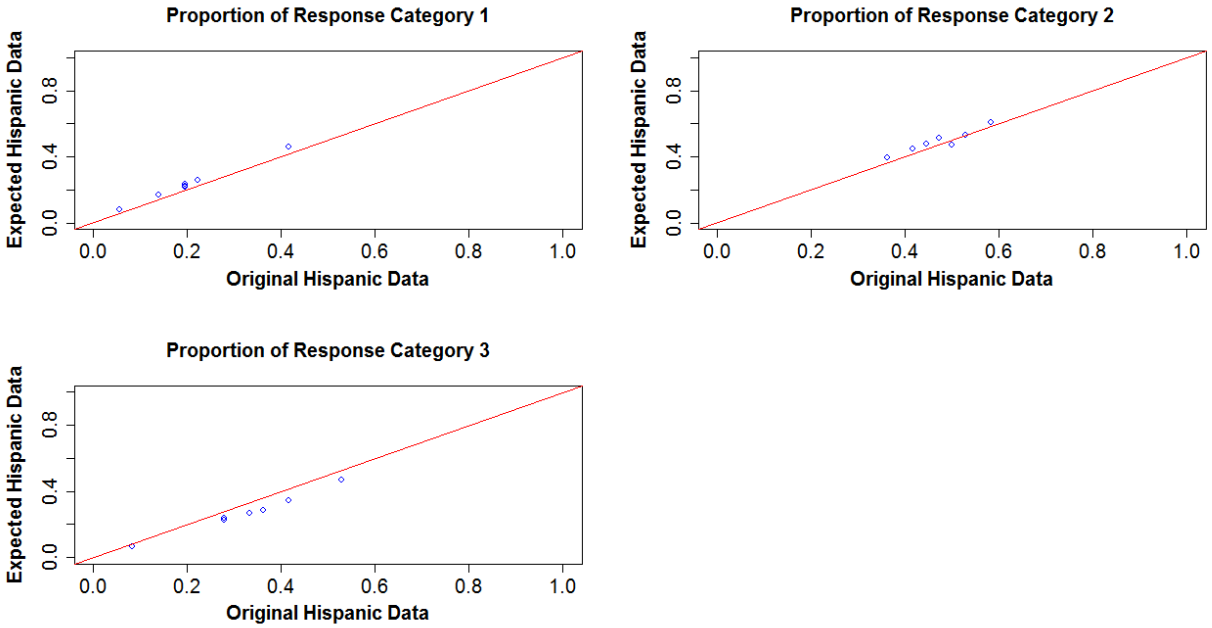


Figure S3.1. Comparison between original vs. expected data for the proportion of Hispanic participants selecting each response option across all seven items.